Environmental Modelling & Software 101 (2018) 289–300

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Uncertainty analysis of a semi-distributed hydrologic model based on a Gaussian Process emulator

Jing Yang [a], [b], [*], Anthony Jakeman [c], Gonghuan Fang [b], Xi Chen [b]

[a] *National Institute of Water and Atmospheric Research, Christchurch, New Zealand*
[b] *State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Xinjiang 830011, China*
[c] *Fenner School of Environment and Society, Australian National University, Australia*

ARTICLE INFO

Article history:
Received 9 February 2017
Received in revised form
25 November 2017
Accepted 29 November 2017
Available online 10 January 2018

Keywords:
Uncertainty analysis
Hydrologic modelling
Sensitivity analysis
Gaussian process emulator
GLUE

ABSTRACT

Despite various criticisms of GLUE (Generalized Likelihood Uncertainty Estimation), it is still a widely-used uncertainty analysis technique in hydrologic modelling that can give an appreciation of the level and sources of uncertainty. We introduce an augmented GLUE approach based on a Gaussian Process (GP) emulator, involving GP to conduct a Bayesian sensitivity analysis to narrow down the influential factor space, and then performing a standard GLUE uncertainty analysis. This approach is demonstrated for a SWAT (Soil and Water Assessment Tool) application in a watershed in China using a calibration and two validation periods. Results show: 1) the augmented approach led to the screening out of 14–18 unimportant factors, effectively narrowing factor space; 2) compared to the more standard GLUE, it substantially improved the sampling efficiency, and located the optimal factor region at lower computational cost. This approach can be used for other uncertainty analysis techniques in hydrologic and non-hydrologic models.

1. Introduction

Distributed hydrological modelling is a way to understand site-specific hydrology and support operational management, planning and decision making in water resources under various scenarios (e.g., water use, land use change, or climate change) (Arnold et al., 2015). Prior to its application, models generally go through a so-called calibration process. Although the hydrologic and other environmental modeling communities have generally promoted the concept and desirability of uncertainty analysis (UA) (e.g. Beven and Freer, 2001; Todini, 2007; Jakeman et al., 2006), there is still a need for its more widespread general practice, especially for distributed hydrologic modelling (Muleta and Nicklow, 2005; Yen et al., 2015). For example, most applications are still based on reporting a single optimum parameter set (Fang et al., 2015; Liu et al., 2011; Luo et al., 2012). The reasons are multifold; for example, some uncertainty analysis techniques are difficult to apply (e.g. the need for testing statistical assumptions in Bayesian inference; see Yang et al., 2008 for a discussion), or only one parameter set is intended or sought for decision making (Song et al., 2015). For complex environmental models, as occur in distributed hydrologic modelling, a restriction would often be the number of model runs required for the UA, which can be a burden, even with the ongoing advances in information technology (e.g. increase of CPU speed and parallel computation technology).

Various quantitative UA techniques have been developed or applied in the literature (Matott et al., 2009) and there are now many societies and journals that promote UA. In the hydrologic modelling literature, such techniques include Generalized Likelihood Uncertainty Estimation or GLUE for short (Beven and Binley, 1992), SUFI (Abbaspour et al., 2007), first-order approximation (Vrugt and Bouten, 2002), and Bayesian inference (Kuczera and Parent, 1998; Kavetski et al., 2006; Yang et al., 2007). Among all these techniques, GLUE is still by far the most widely applied technique in hydrology (Shen et al., 2012; Stedinger et al., 2008) due to its simplicity and practicality, though it has been criticized for several reasons including its informal statistical basis, sampling inefficiency and flat response surface (Mantovan and Todini, 2006; Yang et al., 2008; Beven and Binley, 2014). On the other hand, it can be argued that GLUE warrants use as a guide at least to appreciating the level of various sources of uncertainty. When applying GLUE, however, one might face a substantial computational burden as

[*] Corresponding author. National Institute of Water and Atmospheric Research, 10 Kyle Street, Christchurch 8011, New Zealand.
E-mail address: jing.yang@niwa.co.nz (J. Yang).

https://doi.org/10.1016/j.envsoft.2017.11.037
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

captured in Beven and Binley (2014; Page 5905) who state "it remains an issue, either because of a model that is particularly slow to run so that it is still not possible to sample sufficient realizations or because of a high number of parameter dimensions". This often arises in spite of suggestions (e.g. McMillan and Clark, 2009) to seek increases in the efficiency of finding behavioral models. It has been noted that the sample efficiency (i.e., number of behavioral sets over number of sampling sets) of some applications can be lower than 10^{-4} (Iorgulescu et al., 2005, 2007; Yang et al., 2008).

Over recent decades, emulators (or in some literature metamodels or surrogate models) have been widely applied as a surrogate to deterministic models, partly to overcome the high computational cost of the latter. But certain types of surrogates like Gaussian Processes can also be used to assess properties of the model response surface (e.g. see Asher et al., 2015 for a review in the groundwater domain). These emulators include polynomial regression (Jones, 2001), multivariate adaptive regression splines (Friedman, 1991), radial basis functions (Dyn et al., 1986), polynomials chaos (Wiener, 1938; Xiu and Karniadakis, 2002) and Gaussian Processes (Kennedy and O'Hagan, 2001; Sacks et al., 1989). Most of these applications (especially in hydrologic modelling) have been for optimization (Emmerich et al., 2006; Jones, 2001) and global sensitivity analysis (Oakley and O'Hagan, 2004; Ratto et al., 2007).

In this paper, we propose an emulation-augmented GLUE, using Gaussian Process (GP) emulation of the original model, to help conduct uncertainty analysis. This GLUE-GP uses global and local sensitivity analysis arising as a natural byproduct of the GP emulation to screen out unimportant parameters and reduce the ranges of more sensitive ones, implemented in the software GEM-SA (www.tonyohagan.co.uk/academic/GEM/), before application of GLUE. The GP also allows improvement in the sampling efficiency and location of the optimal region for GLUE sampling at a much lower computational cost than the standard GLUE. The GLUE-GP method is thus akin to GLUE in the sense that it is an augmentation that applies a GLUE procedure but only to those factors and their ranges as informed by the initial GP emulation and its inherent sensitivity analysis. Thus it is an approximation of GLUE whose differences are numerically investigated here in both calibration and validation modes. This approach is demonstrated on the semi-distributed hydrologic model SWAT (Soil and Water Assessment Tool; Arnold et al., 1998) with an application to the Kaidu River Basin in Xinjiang, China, which is an important water source for human activity and ecological function in the oasis downstream.

The remainder of this paper is structured as follows: section 2 gives a brief introduction to GLUE and the GP emulator, and then focuses on the proposed emulation-based GLUE approach (GLUE-GP) and case study; section 3 introduces the SWAT model and case study area; section 4 presents and discusses results; and finally conclusions are summarized in section 5.

2. Methodology

A large class of hydrologic and environmental models can be formulated as $y = f(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ is a vector of m factors and y is either scalar or vector model output (e.g., flow rate time series) or objective function (e.g., root mean square error between simulated and observed flows), and the notation x_i^j to indicate the jth realization of the ith factor of $\boldsymbol{x}$. Here, we distinguish factors from model parameters in that a factor could be a model parameter or a modification to a distributed parameter either in a relative way or with a replacement to their initial values (examples are given in Table 1 and section 3.2). Thus we use the terminology factors instead of model parameters for the GP and uncertainty analyses below.

2.1. GLUE

GLUE (Beven and Binley, 1992, 2014) is an uncertainty analysis technique inspired by the regional sensitivity analysis of Hornberger and Spear (1981). In contrast to assuming that there is a single "optimal" factor set for a model, it is based on the concept of "equifinality" in which different "behavioral" factor sets lead to similarly good model results in some sense. It recognizes that most environmental models used for prediction are non-identifiable due largely, but not only, to the over-parameterised structure of the model (see Shin et al., 2015 for an overview of methods to check structural identifiability). Fig. 1(a) shows a typical procedure for uncertainty analysis based on GLUE.

When applying GLUE, one needs to define an objective function *L(.)* (or "generalized likelihood measure"), and a given threshold value which is used to assess if a sampled factor set is "behavioral" or "non-behavioral" through a comparison: if the corresponding "likelihood measure" is better or worse than the given threshold value. Each behavioral factor is then given a "likelihood weight" according to:

$$w_i = \frac{L(x^i)}{\sum_{k=1}^{N} L(x^k)} \tag{1}$$

where $L(x^i)$ is the objective function value of factor set x^i, and N is the number of behavioral factor sets from $\mathrm{N_T}$ total samples. Then the model predictive uncertainty is described as a prediction band from the cumulative distribution of the model output realized from the weighted behavioral factor sets.

Based on these behavioral and non-behavioral factor sets, factor sensitivity can also be studied with Regionalized Sensitivity Analysis (RSA; Spear and Hornberger, 1980) which inspired GLUE but is not generally part of GLUE. The idea of RSA is: if the distributions of a factor in the behavioral and non-behavioral factor sets are dissimilar then this factor is considered influential. In practice, this is performed with the Kolmogorov-Smirnov test to obtain a distance measure (D) which is the maximum distance between the two empirical cumulative distributions (i.e. behavioral and non-behavioral).

2.2. Gaussian Process emulator

The emulator we invoke is the Gaussian Process emulator, although other emulators such as those based on Polynomial Chaos (Wiener, 1938; Xiu and Karniadakis, 2002) have similar advantages, mainly sampling efficiency improvements and global and local sensitivity measures as byproducts of the emulation (largely because such emulators and their response surfaces are continuous functions which can be differentiated). The idea is to construct a simpler and computationally efficient model as a surrogate for the complicated (more physically based) and less computationally efficient model. When applying a GP emulator to a hydrologic model $y = f(\boldsymbol{x})$, it approximates $f(\boldsymbol{x})$ as a Gaussian process (Kennedy and O'Hagan, 2001)

$$\widehat{f}(\boldsymbol{x}) = m(\boldsymbol{x}) + e(\boldsymbol{x}) = h(\boldsymbol{x})^T\boldsymbol{\beta} + e(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{x}$ *is* a vector of factors, $m(\boldsymbol{x})$ is the mean function, $h(\boldsymbol{x})$ a known regression function, and $e(\boldsymbol{x})$ a zero mean Gaussian process with correlation given by $cov[f(\boldsymbol{x}), f(\boldsymbol{x}')] = \sigma^2 c(\boldsymbol{x}, \boldsymbol{x}')$. In this study $c(\boldsymbol{x}, \boldsymbol{x}')$ is a correlation function between two points $\boldsymbol{x}$ and $\boldsymbol{x}'$ that takes the form $\exp[-(\boldsymbol{x} - \boldsymbol{x}')\boldsymbol{\theta}(\boldsymbol{x} - \boldsymbol{x}')]$.

Table 1
Selected factors, their initial ranges here and underlying SWAT parameters in Arnold et al. (2012) ('*v*__' or '*r*__' in second column means a replacement or a relative change to the initial factor values).

No.	Factor	Range	Underlying SWAT parameter
1	*v__Tlaps*	[-10, −4]	*Tlaps:* Temperature lapse rate (°C km^{-1})
2	*v__Alpha_bf*	[0, 1]	*Alpha_bf:* Baseflow alpha factor
3	*v__Plaps*	[100, 200]	*Plaps:* Precipitation lapse rate (mm km^{-1})
4	*v__Gwqmn*	[0, 1000]	*Gwqmn:* Threshold water level in shallow aquifer for baseflow (mm)
5	*r__Sol_k*	[-0.5, 0.5]	*Sol_kl*: Saturated hydraulic conductivity (mm h^{-1})
6	*v__Gw_delay*	[0, 500]	*Gw_delay*: Groundwater delay time (day)
7	*v__Esco*	[0, 1]	*Esco*: Soil evaporation compensation factor (−)
8	*r__Slsubbsn*	[-0.3, 0.3]	*Slsubbsn*: Average slope length (m)
9	*v__Ch_k2*	[0, 500]	*Ch_k2*: Effective hydraulic conductivity of main channel (mm h^{-1})
10	*r__Sol_awc*	[-0.5, 0.5]	*Sol_awc*: Available water capacity of the soil layer (−)
11	*r__CN2*	[-0.15, 0.15]	*CN2*: SCS runoff curve number for moisture condition
12	*v__Smfmx*	[-0, 10]	*Smfmx*: Snowmelt factor on June 21 (mm °C $^{-1}$·d^{-1})
13	*r__Sol_z*	[-0.5, 0.5]	*Sol_z*: Depth from soil surface to bottom of layer (mm)
14	*v__Gw_revap*	[-0.02, 0.2]	*Gw_revap*: Groundwater "revap" coefficient
15	*v__Surlag*	[0, 24]	*Surlag*: Surface runoff lag time (day)
16	*v__Revapmn*	[0, 500]	*Revapmn*: Threshold depth of water in shallow aquifer for revap (mm)
17	*r__Slope*	[-0.1, 0.1]	*Slope*: Average slope steepness (−)
18	*v__Ch_k1*	[0, 300]	*Ch_k1*: Effective hydraulic conductivity of tributary channel (mm h^{-1}).
19	*v__Smfmn*	[0, 10]	*Smfmn*: Snowmelt factor on Dec. 21 (mm °C^{-1}·d^{-1})
20	*v__Epco*	[0, 1]	*Epco*: Plant uptake compensation factor
21	*v__Ch_n2*	[0, 0.3]	*Ch_n2*: Manning's "n" for main channel (−)
22	*r__OV_N*	[-0.5, 0.5]	*OV_N*: Manning's "n" for overland flow (−)
23	*r__Sol_alb*	[-0.2, 0.2]	*Sol_alb*: Moist soil albedo (−)
24	*v__Sftmp*	[-1, 1]	*Sftmp*: Snowfall temperature (°C)
25	*v__Smtmp*	[-1, 1]	*Smtmp*: Snow melt base temperature(°C)

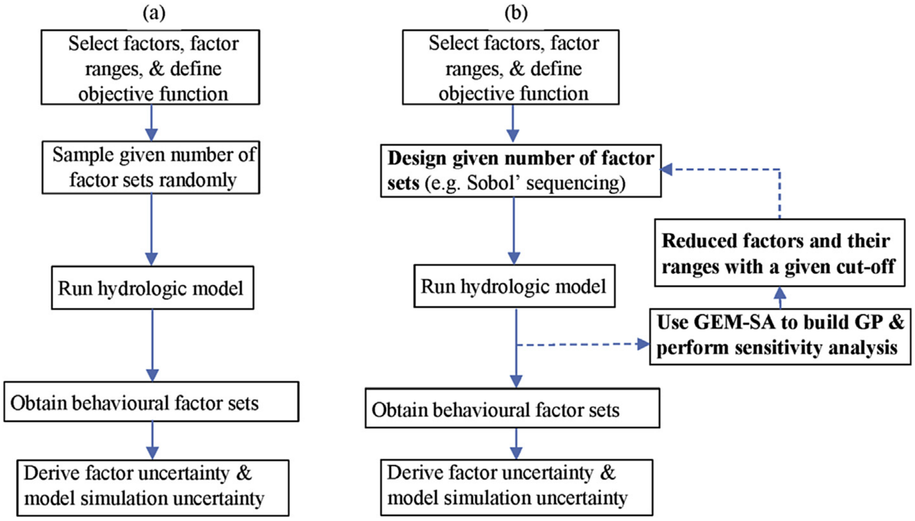
(a)
Select factors, factor ranges, & define objective function
Sample given number of factor sets randomly
Run hydrologic model
Obtain behavioural factor sets
Derive factor uncertainty & model simulation uncertainty
(b)
Select factors, factor ranges, & define objective function
Design given number of factor sets (e.g. Sobol' sequencing)
Reduced factors and their ranges with a given cut-off
Run hydrologic model
Use GEM-SA to build GP & perform sensitivity analysis
Obtain behavioural factor sets
Derive factor uncertainty & model simulation uncertainty

Fig. 1. Uncertainty analysis procedure for (a) GLUE and (b) GLUE-GP where dashed lines indicate the main GP process. GEM-SA is the software used for GP construction and Bayesian sensitivity analysis.

The quantities $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\theta}$ are emulator parameters (hyperparameters). The key assumption is that $f(\boldsymbol{x})$ is a smooth function (no discontinuities, jumps or sharp changes), which allows us to predict $f(\boldsymbol{x}')$ if $\boldsymbol{x}'$ is close to $\boldsymbol{x}$ and $f(\boldsymbol{x})$ is known. Emulator hyperparameters (i.e., $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\theta}$) are estimated based on n design data points $[(x^i, y^i), i = 1,..,n]$ as $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\theta}}$ (Bastos and O'Hagan, 2009). Based on Bayesian inference,

$$\frac{f(\boldsymbol{x})-m^*(\boldsymbol{x})}{\hat{\sigma}\sqrt{c^*(\boldsymbol{x},\boldsymbol{x}')}} \sim t_{n-m} \tag{3}$$

where

$$m^*(\boldsymbol{x}) = h(\boldsymbol{x})^T\hat{\boldsymbol{\beta}} + t(\boldsymbol{x})^T A^{-1}\left(y^D - \boldsymbol{H}\hat{\beta}\right)$$

$$c^*\left(\boldsymbol{x},\boldsymbol{x}'\right) = c\left(\boldsymbol{x},x'\right) - t(\boldsymbol{x})^T\boldsymbol{A}^{-1}t(\boldsymbol{x}) + \left[h(\boldsymbol{x})^T - t(\boldsymbol{x})^T\boldsymbol{A}^{-1}\boldsymbol{H}\right]\left(\boldsymbol{H}^T\boldsymbol{A}^{-1}\boldsymbol{H}\right)^{-1}\left[h(\boldsymbol{x})^T - t(\boldsymbol{x})^T\boldsymbol{A}^{-1}\boldsymbol{H}\right]^T$$

$$t(\boldsymbol{x}) = \left[c\left(\boldsymbol{x},\boldsymbol{x}^1\right),\ldots,c(\boldsymbol{x},\boldsymbol{x}^n)\right]$$

$$\boldsymbol{H}^T = \left[h\left(\boldsymbol{x}^1\right)^T,\ldots,\ h(\boldsymbol{x}^n)^T\right]$$

$$\boldsymbol{y}^D = \left[f\left(\boldsymbol{x}^1\right),\ldots,\ f(\boldsymbol{x}^n)\right]^T$$

$$\boldsymbol{A} = \begin{bmatrix} c\left(\boldsymbol{x}^1, \boldsymbol{x}^1\right) & \cdots & c\left(\boldsymbol{x}^1, \boldsymbol{x}^n\right) \\ \vdots & \ddots & \vdots \\ c\left(\boldsymbol{x}^n, \boldsymbol{x}^1\right) & \cdots & c(\boldsymbol{x}^n, \boldsymbol{x}^n) \end{bmatrix}$$

and t_{n-m} denotes a Student's t-distribution with n − m degrees of freedom.

One advantage of a GP emulator is that it not only gives the prediction (i.e. as a surrogate to the original model) but also the underlying sensitivity of factors resulting from the emulation process. Since these types of integrals are easy to be solved analytically, it is useful for statistical analysis of the emulated response surface, for example variance based sensitivity analysis as explained in the section below.

2.3. Sensitivity analysis

Sensitivity Analysis (SA) techniques can be categorized broadly into local SA and global SA based on the factor space of interest (Saltelli et al., 2008; Norton, 2015; Yang, 2011), with global SA techniques presently being more popular. Among the different global SA techniques, variance based SA is considered the most reliable (Saltelli, 2002). It is based on ANOVA (Analysis of Variance) decomposition, i.e., the original model $y = f(x)$ can be decomposed into m main effects and m interactions:

$$y = f(\boldsymbol{x}) = g_0 + \sum_{i=1}^{m} g_i(x_i) + \sum_{i<j}^{m} g_{ij}(x_i, x_j) + \ldots + g_{1,2,\ldots,m}(\boldsymbol{x}) \qquad (4)$$

where

$$g_0 = E(Y)$$

$$g_i(x_i) = E(Y|x_i) - g_0$$

$$g_{ij}(x_i, x_j) = E(Y|x_i, x_j) - g_i(x_i) - g_j(x_j) - g_0$$

and so on. The quantities $g_i(x_i)$ and $g_{ij}(x_i, x_j)$ are called the main effect and first order interaction, respectively. The sensitivities of the m main effects and m interactions sum to unity, i.e.

$$\sum_{i=1}^{m} S_i + \sum_{i<j}^{m} S_{ij} + \ldots + S_{1,2,..,m} \; = 1 \qquad (5)$$

where $S_i = \frac{V[E((Y|x_i))]}{V(Y)}$ is the main effect index of factor i, $S_{ij} = \frac{V[E(Y|x_i, x_j)]}{V(Y)}$ is the first order interaction index between factor x_i and factor x_j, and so on. The notation V denotes variance.

Among these indices in Eq. (5), S_i represents the average output variance reduction that can be achieved when x_i is fixed. Additionally, the total effect index $S_{Ti} = 1 - \frac{V[E(Y|x_{-i})]}{V(Y)}$, where $\boldsymbol{x}_{-\boldsymbol{i}}$ is the subvector of $\boldsymbol{x}$ containing all elements except x_i, denotes the average output variance that would remain as long as x_i stays unknown. S_{Ti} is used to screen out unimportant factors, and S_i is used to determine the important factor to be fixed in the calibration process (Ratto et al., 2007). Traditional methods to estimate these indices from multiple samplings of the model factors include the Sobol' method (Saltelli, 2002; Sobol', 1990) and the extended FAST method (Saltelli et al., 1999). However these methods are computationally expensive, and emulator based methods have been proposed to reduce the required sampling and associated model runs (e.g. Oakley and O'Hagan, 2004; Ratto et al., 2007).

2.4. Proposed augmented GLUE procedure (GLUE-GP)

The proposed uncertainty analysis procedure for combining GLUE and the GP emulation is shown in Fig. 1 (b). GP construction and sensitivity analysis are performed with the software GEM-SA. Compared to a typical GLUE in Fig. 1 (a), the main procedure is:

1) Based on a design dataset of n sampled points [(x^i, y^i), i = 1,..,n], i.e., designed factor sets and corresponding n model simulation results, approximate $y = f(\boldsymbol{x})$ with the GP emulator by estimating its hyperparameters using the software GEM_SA (www.tonyohagan.co.uk/academic/GEM/).
2) As a product of GEM-SA, a Bayesian sensitivity analysis (Oakley and O'Hagan, 2004) is used to decrease model dimensions (i.e. number of factors) based on the total effect index S_{Ti}; and to narrow factor ranges based on S_i, S_{ij}, and the expectations of $E(y|x_i)$ and $E(y|x_i)$ with a given cut-off. The cut-off, which is a different variable to the threshold in GLUE to determine behavioral and non-behavioral factor sets, is used to narrow factor ranges. Though it is subjective, it should be within the range of minimum and maximum of all "$E(y|x_i)$"s, and based on our experience the average of minimum and maximum all "$E(y|x_i)$"s is a good start, though one can always look at values beyond our illustrative choices.
3) N_M points are sampled from the reduced and range-narrowed factors, and then the corresponding N_M model simulation results are calculated.
4) Uncertainty analysis is then undertaken similarly to the standard GLUE. It is noted that the final behavioral factor sets include all behavioral factor sets from n design points for GP construction at step 1 and N_M points at step 3.

To compare the performance of this GLUE-GP procedure with the basic GLUE on the original model, three measures are computed. The *e-factor* (the percentage of behavioral sets identified), *p-factor* [the percentage of observations bracketed by 95% Prediction Uncertainty (95PPU)] and *r-factor* (relative average width of 95PPU), are calculated according to:

$$e-factor = \frac{N_{bs}}{N_s} \qquad (6)$$

$$r-factor = \frac{\sum_{t_i}^{T}\left(y^M_{t_i,97.5\%} - y^M_{t_i,2.5\%}\right)}{T\ \sigma_{fobs}} \qquad (7)$$

where N_{bs} and N_s are the number of behavioral sets and number of total samples, $y^M_{t_i,97.5\%}$ and $y^M_{t_i,2.5\%}$ (t_i = 1, …, T) represent the upper part and lower part of the 95 PPU of the simulated time series (in the case of our model application the series is streamflow), σ_{fobs} denotes the standard deviation of the measured time series data, T is the number of model outputs (in our case, streamflow), and M indices "modelled".

The *e-factor* quantity, ranging from 0 to 100%, is used to quantify the sampling efficiency, whereas the *p-factor* and *r-factor* quantify the goodness of prediction uncertainty, i.e., closeness of the *p-factor* to 100% (i.e., all observations bracketed by the prediction uncertainty) and *r-factor* to 0 (i.e., achievement of rather small uncertainty band). These three metrics have been invoked in UA using SUFI (Abbaspour et al., 2007) and in a UA technique comparison study (e.g., Yang et al., 2008).

2.5. Sampling techniques

To apply GLUE and GLUE-GP, one needs to sample from prior factor distributions. There are various sampling techniques that have been applied in the literature. The most frequently used include simple random sampling (Monte Carlo), Markov chain Monte Carlo (MCMC), importance sampling, Latin hypercube sampling (LHS), and Sobol' sequencing. Simple random sampling, often used with GLUE, is an entirely random sampling technique. However it “tends to undersample small but possibly very important areas unless the sample size is significantly increased” (Congalton, 1991). LHS uses a stratified sampling scheme to improve on the coverage of factor space (McKay et al., 1979), and Sobol' sequencing belongs to the family of quasi-random low-discrepancy sequences (Sobol' et al., 2011). It uses a base of two to form successively finer uniform partitions of the unit interval and then reorders the coordinates in each dimension (see more details in Sobol' et al., 2011). Some studies show that Sobol' sequences (e.g. Kucherenko et al., 2016; Singhee and Rutenbar, 2009) and LHS (e.g. Aistleitner et al., 2012) have faster convergence than simple random sampling.

In this study, we use the Sobol' sequence for sampling as it typically has a good coverage of factor space and is quite often used to compute Sobol' indices (Sobol’ 2001). However, we encourage similar studies on other sampling techniques.

3. SWAT model and study area

3.1. SWAT model

SWAT (Soil and Water Assessment Tool; Arnold et al., 1998) is a semi-distributed hydrologic model and has been widely used to assess the impact of different management practices and climate change on water resource quantity and quality at a watershed scale (e.g., Arnold and Fohrer, 2005; Setegn et al., 2011). It divides the watershed into sub-basins, within which a number of hydrologic response units are generated with a unique combination of land-use, soil type and slope. The sub-basins are linked through a river network for surface water movement. Its climatic input consists of daily precipitation, maximum/minimum air temperature, solar radiation, wind speed and relative humidity. In mountainous regions, SWAT can use elevation bands and precipitation/temperature lapse rates to account for orographic effects on precipitation and temperature. For more details, refer to the SWAT manuals (http://www.brc.tamus.edu/).

3.2. Study area and model setup

The Kaidu River Basin, with a drainage area of 18,634 km^2 above the Dashankou hydrological station, is located in the Tianshan Mountains in northwest China (Fig. 2). It is characterized as temperate continental with alpine climate. As one of the headwaters of the Tarim River, it provides water resources for agricultural activity and the ecological environment of the oases in the lower reaches (Fang et al., 2015). The altitude ranges from 1342 m to 4796 m above sea level (a.s.l.) with an average elevation of 2995 m, so the orographic effect on precipitation and temperature is significant and is represented here. Observed flow data are available either at daily or monthly scale from 1980 to 2010, and were used here.

Daily observed meteorological data, including precipitation, maximum/minimum temperature, wind speed and relative humidity of two meteorological stations (Bayanbulak and Baluntai locations are denoted by asterisks in Fig. 2, and are 2458 m and 1740 m a.s.l. respectively), are from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/). The mean annual maximum and minimum temperature at the Bayanbulak meteorological station are 3.1 °C and −10.6 °C and mean annual precipitation is 267 mm; generally, precipitation falls as rain from May to September and as snow from October to April of the next year. The observed streamflow data at the Dashankou hydrologic station (the triangle in Fig. 2) are from the Xinjiang Tarim River basin Management Bureau. The average daily flow is around 110 m^3 s^{-1} (equivalent to 185 mm of runoff per year), ranging from 15 m^3 s^{-1} to 973 m^3 s^{-1}. For more details, refer to Fang et al. (2015).

Previously Fang et al. (2015) set up a SWAT model in this watershed with available DEM (digital elevation model), land use, soil, and observed climate data, and studied the contribution of meteorological inputs to hydrologic modelling results and the impact of climate change on the water resources based on a single optimized factor set. The purpose of using the same case is to test our proposed uncertainty analysis approach for its advantages over a typical GLUE, accepting that uncertainties could be reduced and characterized by either approach if better climate and streamflow information were available.

Table 1 lists the initial 25 selected factors for the calibration, which are assumed to be uniformly distributed as in Fang et al. (2015). Factors, distinct from model parameters, can be modifications to model parameters either in a relative way or a replacement; for example in Table 1 the factor *r__Sol_k* is a relative change to the distributed parameter *Sol_k* (saturated hydraulic conductivity) while *v__Tlaps* is a replacement of initial values of *Tlaps* (Temperature lapse rate) in order to lower the calibration dimension. Standardized root mean squared error (*SRMSE*) is invoked in the paper for demonstration as the so-called likelihood measure (objective function):

$$SRMSE = \frac{\sqrt{\sum_1^T \left(y_{t_i}^M - y_{t_i}\right)^2 \Big/ T}}{\sigma_{fobs}} \tag{8}$$

where y_{t_i} is the observed outflow at time t_i. *SRMSE* can be related to the widely used Nash-Sutcliffe coefficient (*NS*) through $NS = 1 - SRMSE^2 \cdot \frac{T}{T-1} \approx 1 - SRMSE^2$. A value of 0.55 for example for *SRMSE* is equivalent to an *NS* value of 0.70.

In this study, GLUE and GLUE-GP were first applied to a calibration period from 1986 to 1989 with daily flow data, and then to two validation periods, one from 1990 to 2002 with daily flow data, and the other from 2003 to 2010 with monthly flow data, which are the same as the calibration and validation periods in Fang et al. (2015).

4. Results

Based on an initial selected 25 factors (Table 1), which is a fairly standard level of parameterisation for a SWAT model, and the likelihood measure *SRMSE*, 600 Sobol' sequence samples were generated as design data and a corresponding 600 daily model simulations were obtained. Of these samples, the first 300 were used to train the emulator and the remaining 300 for so-called validation. The emulation result is shown in Fig. 3. Black dots signify training data and grey dots validation points. It is not surprising that training data are on the 1:1 line, which is characteristic of a GP emulator. Data for validation were scattered around the 1:1 line with an *NS* value of 0.82 for the emulated *SRMSE* fit to the original modelled *SRMSE*. A similar result was obtained through a cross-validation (using the last 300 design samples to train the emulator and first 300 samples for validation).

As part of the GP construction, a Bayesian sensitivity analysis was undertaken for these 25 factors with GEM-SA software, which uses the same samples as were generated for the GP construction. The results are captured in Table 2 and Fig. 4. In Table 2, of the main

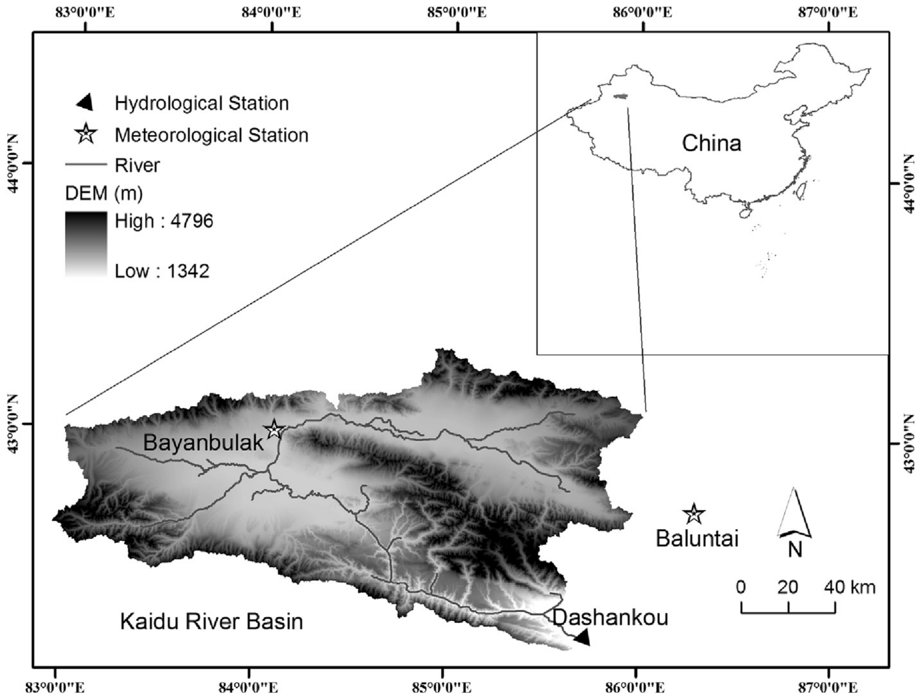

Fig. 2. The location, topography and river system of the Kaidu River Basin.

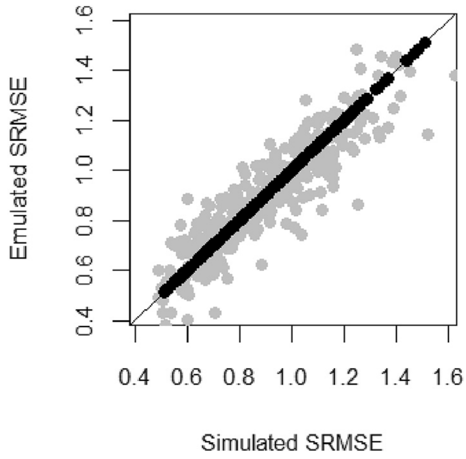

Fig. 3. Performance of the GP emulator for training data (black) and validation data (grey).

Table 2
Main effect (S_i) and total effect (S_{Ti}) indices based on Bayesian sensitivity analysis, and distance measure (D) based on Regional Sensitivity Analysis (RSA).

Factor	S_i	S_{Ti}	D	Factor	S_i	S_{Ti}	D
v__Tlaps	38.95	54.17	0.67	*v__Gw_revap*	0.36	0.84	0.08
v__Alpha_bf	15.83	28.16	0.49	*v__Surlag*	0.22	0.53	0.12
v__Plaps	4.64	10.02	0.22	*v__Revapmn*	0.07	2.26	0.08
v__Gwqmn	3.88	9.57	0.16	*r__Slope*	0.07	0.07	0.08
r__Sol_k	1.96	3.9	0.18	*v__Ch_k1*	0.79	2.03	0.14
v__Gw_delay	2.1	14.94	0.29	*v__Smfmn*	0.2	0.36	0.04
v__Esco	0.84	4.53	0.06	*v__Epco*	0.2	0.32	0.06
r__Slsubbsn	0.8	2.35	0.05	*v__Ch_n2*	0.08	0.09	0.09
v__Ch_k2	0.06	0.06	0.09	*r__OV_N*	0.05	0.05	0.07
r__Sol_awc	1.22	2.57	0.05	*r__Sol_alb*	0.06	0.51	0.04
r__CN2	0.04	0.06	0.08	*v__Sftmp*	0	0.01	0.02
v__Smfmx	0.26	0.93	0.04	*v__Smtmp*	0	0	0.02
r__Sol_z	0.14	0.91	0.08				

effect indices (S_i), the most sensitive factor is *v__Tlaps*, followed by *v__Alpha_bf*, *v__Plaps* and *v__Gwqmn*, while others are not as sensitive. For total effect indices (S_{Ti}), the most sensitive factor is *v__Tlaps*, followed by *v__Alpha_bf*, and then *v__Gw_delay*, *v__Plaps* and *v__Gwqmn*, while others are not as sensitive. The difference between the total effect index of a factor and its main effect index measures the interaction between this factor and other factors. There are interactions for these mentioned factors, especially for *v__Tlaps* and *v__Alpha_bf*. However, all first order interaction indices (i.e. S_{ij}) are around but less than 2%. For example, S_{ij} between *v__Tlaps* and *v__Alpha_bf* is 1.9%, which denotes that first order interactions are low. This result is similar to the result of an approach combining the Morris method and SDP method applied in Fang et al. (2015), and their main effect indices and first order interactions are very close to our results here, except that the approach here includes total effect indices and requires less model runs. Based on this analysis, insensitive factors whose total effect indices are less than 2% were screened out, which led to 11 factors being considered for further analysis. As a comparison, we also undertook a similar analysis for the 7 most sensitive factors whose "S_{ij}"s were over 3.5% (first 7 factors in first column in Table 2). In the following, when not specified, results are based on these 11 most sensitive factors. In contrast, Regional Sensitivity Analysis (RSA) can only detect the first 6 most sensitive factors in Table 2, and identifies "*v__Ch_k1*" and "*v__Surlag*" as sensitive factors that are not sensitive in Bayesian sensitivity analysis. Similar results can also be found in Yang (2011).

Sensitivity analysis can also be used to narrow factor ranges based on posterior expectations of $E(y|x_i)$ (expectation of *SRMSE* conditioned on a given factor x_i in Table 2) and $E(y|x_i, x_j)$. Fig. 4 shows posterior expectations $E(y|x_i)$ of the 11 most sensitive factors and the posterior expectation $E(y|x_i, x_j)$ between *v__Tlaps* and *v__Alpha_bf* (Other "$E(y|x_i, x_j)$"*s are* not shown here as their corresponding S_{ij} are lower than 2%). From the $E(y|x_i)$ plots, lower *SRMSE* values are concentrated in the left side of the initial range for

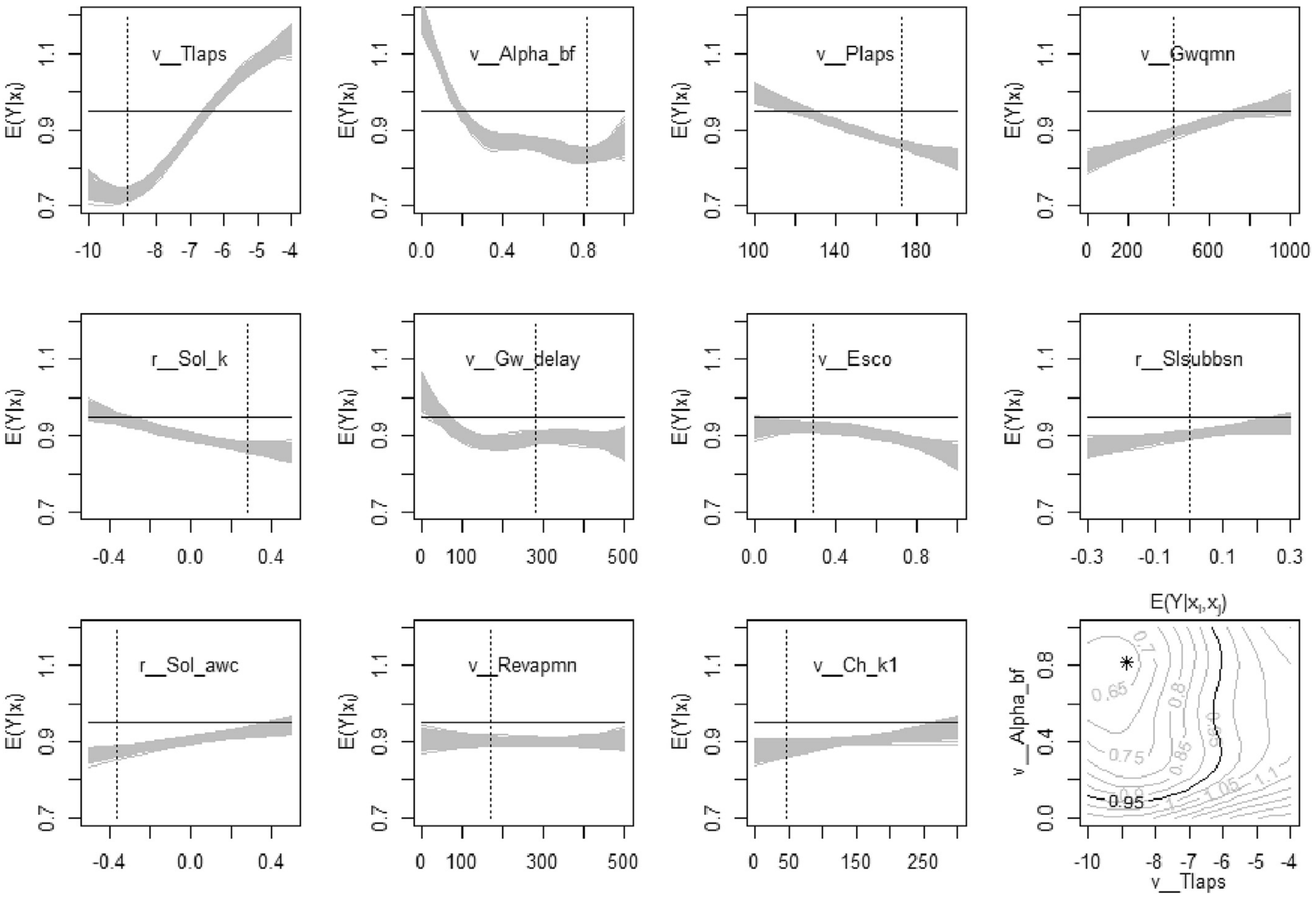

Fig. 4. Posterior expectations of $E(y|x_i)$ (expectation of *SRMSE* conditioned on factor x_i) of the 11 most sensitive factors and $E(y|x_i, x_j)$ between *v__Tlaps* and *v__Alpha_bf* based on GEM-SA. Solid line denotes a cut-off of 0.95 and vertical dotted lines and star are the location of the best factor set.

v__Tlaps, from the middle to the right for *v__Alpha_bf*, *v__Plaps*, and *v__Gw_delay*, while flat for other factors (e.g. *v__Esco*). The posterior expectation of $E(y|x_i, x_j)$ between *v__Tlaps* and *v__Alpha_bf* also shows that low *SRMSE* values are concentrated in the upper corner (lower limit of *v__Tlaps* and upper limit of *v__Alpha_bf*). By setting a cut-off point for the posterior expectation $E(y|x_i)$, the ranges of sensitive factors can be reduced. $E(y|x_i)$ lies in [0.7, 1.25] and this gives a suggested cut-off of 0.975, and so here we chose three cut-offs, i.e., 1.0, 0.95, and 0.9, to examine the effects of departures from 0.975. Table 3 lists the reduced factor ranges for sensitive factors at these three cut-offs. For example, for a cut-off value of 0.95, the new range is [-10, −6.3] for *v__Tlaps*, [0.17, 1] for *v__Alpha_bf*, and [110, 200] for *v__Alpha_bf*, while there was no change in ranges for other factors.

Table 3
Reduced factor ranges for 11-factor GLUE-GP with different cut-offs. Other factors are not listed as their ranges cannot be reduced.

Factor	Cut-off 1.0	Cut-off 0.95	Cut-off 0.9
v__Tlaps	[-10,5.7]	[-10,-6.3]	[-10, −6.7]
v__Alpha_bf	[0.12,1]	[0.17,1]	[0.22,1]
v__Plaps	–	[110,200]	[140,200]
v__Gwqmn	–	–	[0,570]
r__Sol_k	–	–	[-0.12,0.5]
v__Gw_delay	–	–	[70,500]
v__Esco	–	–	[0.05,1]
r__Sol_awc	–	–	[-0.5,0.1]

For the ensuing uncertainty analysis, 2000 Sobol' sequence samples were then generated for each cut-off selection and followed by a corresponding 2000 model runs, which led to 2000 model simulation results for each cut-off. When applying the standard GLUE, 5000 Sobol' sequence samples were generated, followed by a corresponding 5000 model runs, the larger number being required in order to have a better coverage of (the larger) factor space. The uncertainty analysis was conducted separately for GLUE, and each GLUE-GP application (i.e. for each cut-off value of posterior expectation).

Fig. 5 compares the histograms of objective function (i.e. *SRMSE* in Eq. (8)) values based on GLUE (5000 simulations; Top), and GLUE-GP with cut-off values of 1.0, 0.95 and 0.90 (each with 2600 simulations in total, i.e. 600 simulations for GP construction (initial calibration and validation) plus 2000 simulations after range reduction) based on 11 factors and 7 factors, respectively. For GLUE, objective functions are spread out almost uniformly from 0.5 to 1.3, while for GLUE-GP with a cut-off of 1.0, objective function values shift to the lower values and are concentrated around 0.6–0.9 based on 11 factors, and 0.5–0.9 based on 7 factors; and shifts even more to lower values towards the minimum for GLUE-GP with cut-offs 0.95 and 0.90 based on both 11 factors and 7 factors. Compared to the 11-factor based GLUE-GP, 7-factor based GLUE-GP has a higher frequency of *SRMSE* around 0.5.

For illustration, we chose a threshold for *SRMSE* of 0.55

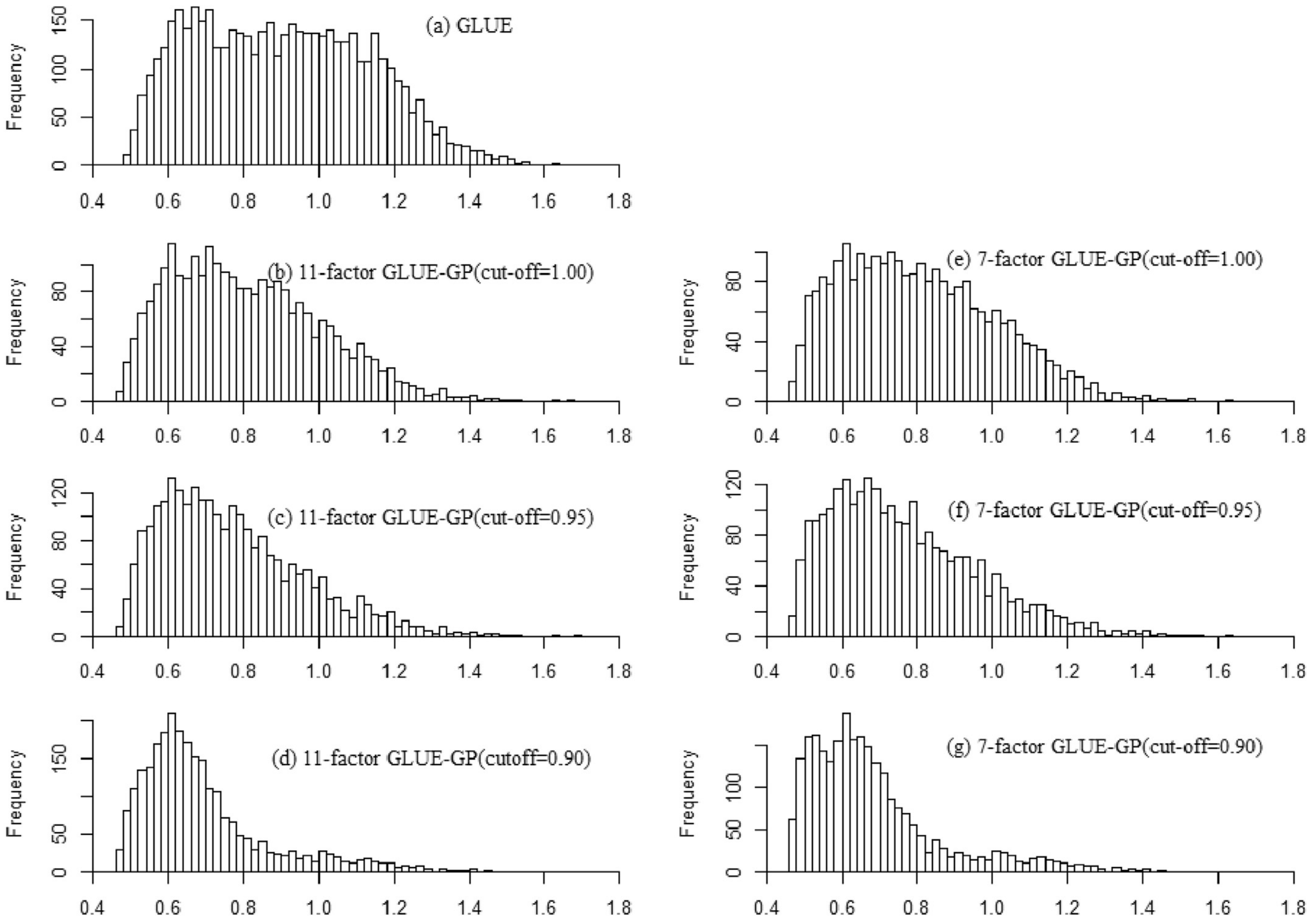

Fig. 5. The histogram of objective functions *SRMSE* for GLUE (a) and 11-factor GLUE-GP with prior expectation cut-offs of 1.00 (b), 0.95 (c), and 0.90 (d) in the calibration period 1986-89. (As a comparison, plots e, f and g are for 7-factor GLUE-GP).

(equivalent to *NS* of 0.70) below which the behavioral sets were selected for uncertainty analysis Fig. 6 plots the 95PPU bands and the best simulations (corresponding to the lowest *SRMSE*) for GLUE, and GLUE-GP with each of the three cut-off values), while Table 4 shows their sampling efficiency, uncertainty statistics and best (smallest) *SRMSE* in both calibration and validation periods. For these 95PPU bands, visually there are no obvious differences between the first three plots. Their relative widths however can be quantified by the *r-factor*, results of which are in Table 4. In the 1986-89 period, as expected, the 95PPU bands of GLUE are the widest (0.83) as it samples from the most expansive prior distributions, followed by GLUE-GP with cut-offs of posterior expectation 1.0 and 0.95 (both 0.78); and that of GLUE-GP with a cut-off of 0.90 is the narrowest (0.74).

Sampling efficiency, given by the *e-factor* or percentage of behavioral sets, is low (3%) for GLUE and significantly increased by GLUE-GP with expected increasing efficiency for decreasing cut-offs (7%, 9%, and 16% for cut-offs of 1.0, 0.95 and 0.9, respectively). Table 4 also summarizes that in the calibration the 95PPU bands of GLUE include 81% of the observed data, the 95PPU bands of the 11-factor based GLUE-GP include similar percentages (81% and 80%) of observed data when cut-offs are 1.0 and 0.95, while this reduces to 77% with a cut-off posterior expectation of 0.90. The 7-factor based GLUE-GP gives similar results to the 11-factor based GLUE-GP for all three cut-offs. The best objective function value from all simulations is lowest and best for the 7-factor based GLUE-GP, followed by the 11-factor based GLUE-GP, while that of GLUE has the worst objective function value.

Given that GLUE-GP applies standard GLUE on a reduced set of factors and factor ranges to GLUE, its computed uncertainties will be somewhat different to that of GLUE. We therefore investigated the level of differences between the two approaches, focusing on the validation periods. In the two validation periods, we found that for both approaches over 95% of the behavioral parameter sets in the calibration period remain behavioral (Table 4), with GLUE-GP values typically higher than GLUE. Similarly, uncertainty results in the calibration period carry over for both approaches to the validation periods, i.e., "*r-factor*"s and "*p-factor*"s of GLUE-GP validations are either close to or slightly lower than those of GLUE, but objective function values of GLUE-GP are lower than those of GLUE.

Fig. 7 indicates the differences through the validation time periods between the two approaches in terms of their 95% prediction uncertainty bands. The main observable differences occur at some of the larger flows, otherwise the bands are qualitatively very similar.

5. Discussion

5.1. Comparison with GLUE

Compared to GLUE, the 11-factor GLUE-GP has an improved sampling efficiency (from 3% to over 7% for all three GLUE-GP

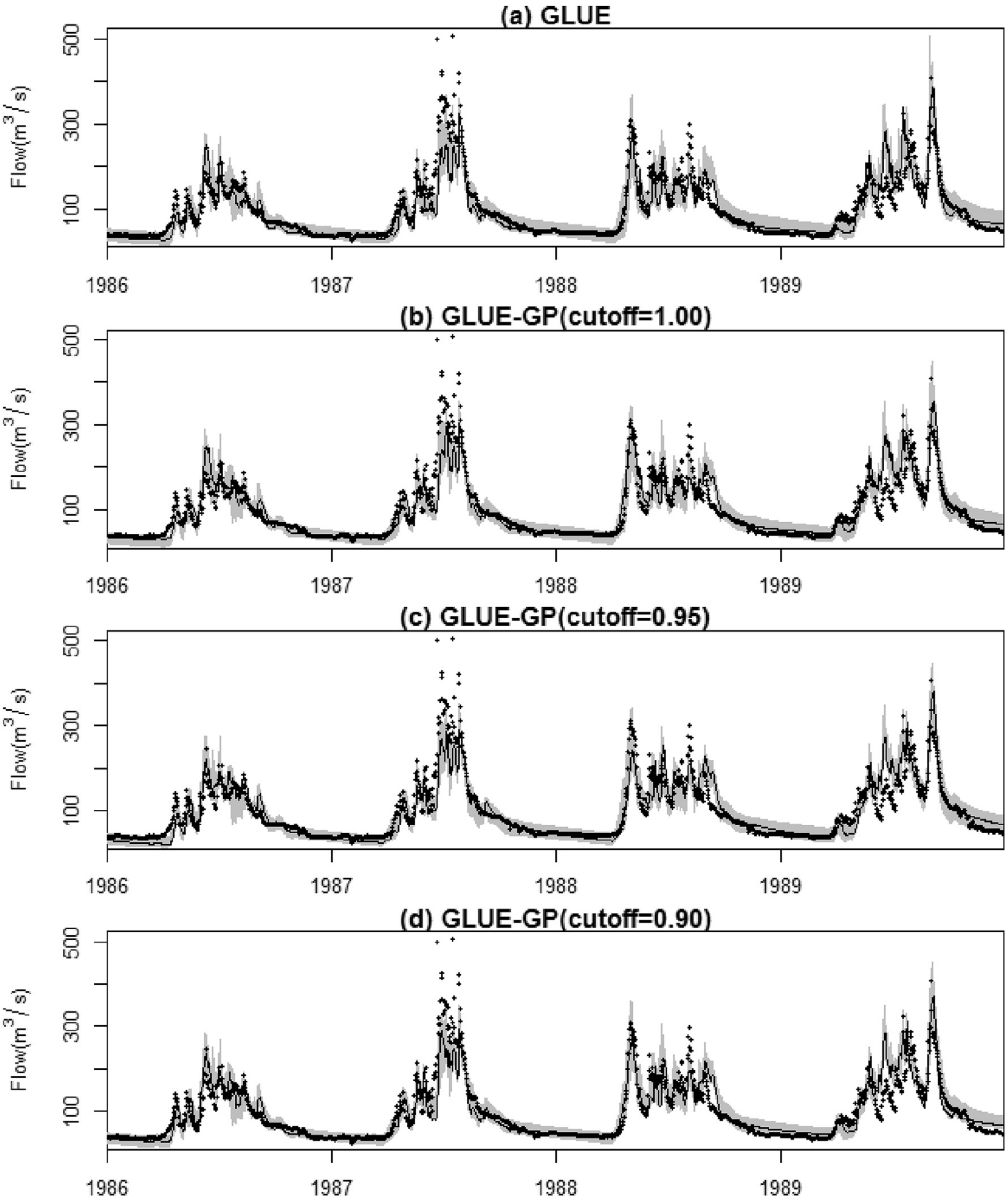

Fig. 6. 95% uncertainty ranges of flow simulations based on a threshold *SRMSE* of 0.55 for GLUE (a) and 11-factor GLUE-GP with cut-offs of 1.00 (b), 0.95 (c), and 0.90 (d) in the calibration period 1986-89.

applications; indeed it is over 9% when only 7 sensitive factors are invoked) over factor space, while obtaining similar uncertainty coverage for posterior expectation cut-offs of 1.0 and 0.95 (80 and 81% versus GLUE's 81%). However one needs to be careful not to enforce cut-offs that are too stringent. In our example, a cut-off of 0.9 increases efficiency to 16% in the 11 sensitive factor case and 22% if only 7 factors are used, but it reduces coverage to 77% in both cases. An acceptable cut-off should always be guided by appropriate selection of the metric(s) for the output of interest (e.g. Bennett et al., 2013) and the accuracy of uncertainty estimates required for it. In many cases, it might be acceptable to have a stringent cut-off that is more efficient computationally yet yields a useful qualitative appreciation of uncertainties. Thus these results are a consequence of the GP technique excluding insensitive factors and narrowing factor ranges through a Bayesian sensitivity analysis that provides Sobol' indices and expectations of $E(y|x_i)$ and $E(y|x_i, x_j)$. GLUE-GP tends to have a stronger focus on optimal factor regions instead of the entire factor space and leads to higher weight on optimal factor regions, which further leads to a better optimal factor estimation (e.g. last row in Table 4).

5.2. Comparison with the Sobol' method

Generally, with the Sobol' method for sensitivity analysis, only

Table 4
Comparison in calibration and validation mode of *e-factor*, *r-factor* and *p-factor* values for GLUE and GLUE-GP with three different cut-offs.

		GLUE	11-factor GLUE-GP			7-factor GLUE-GP		
			cut-off 1.00	cut-off 0.95	cut-off 0.90	cut-off 1.00	cut-off 0.95	cut-off 0.90
Calibration	*e-factor*	3%	7%	9%	16%	9%	12%	22%
1986–1989	*r-factor*	0.83	0.78	0.78	0.75	0.78	0.77	0.74
daily data	*p-factor*	81%	81%	80%	77%	83%	82%	77%
	Best objective	0.485	0.473	0.470	0.465	0.463	0.461	0.462
Validation	*e-factor*[a]	95%	99%	98%	100%	99%	98%	99%
1990–2002	*r-factor*	0.78	0.76	0.77	0.68	0.76	0.75	0.67
daily data	*p-factor*	74%	74%	75%	63%	75%	75%	67%
	Best objective	0.438	0.421	0.422	0.424	0.421	0.422	0.420
Validation	*e-factor*[a]	98%	98%	98%	100%	97%	97%	100%
2003–2010	*r-factor*	0.86	0.90	0.90	0.76	0.88	0.87	0.75
monthly data	*p-factor*	78%	82%	82%	78%	81%	81%	74%
	Best objective	0.354	0.329	0.330	0.318	0.314	0.319	0.317

[a] Percentage of behavioral factor sets in calibration which are still behavioral in the given validation period.

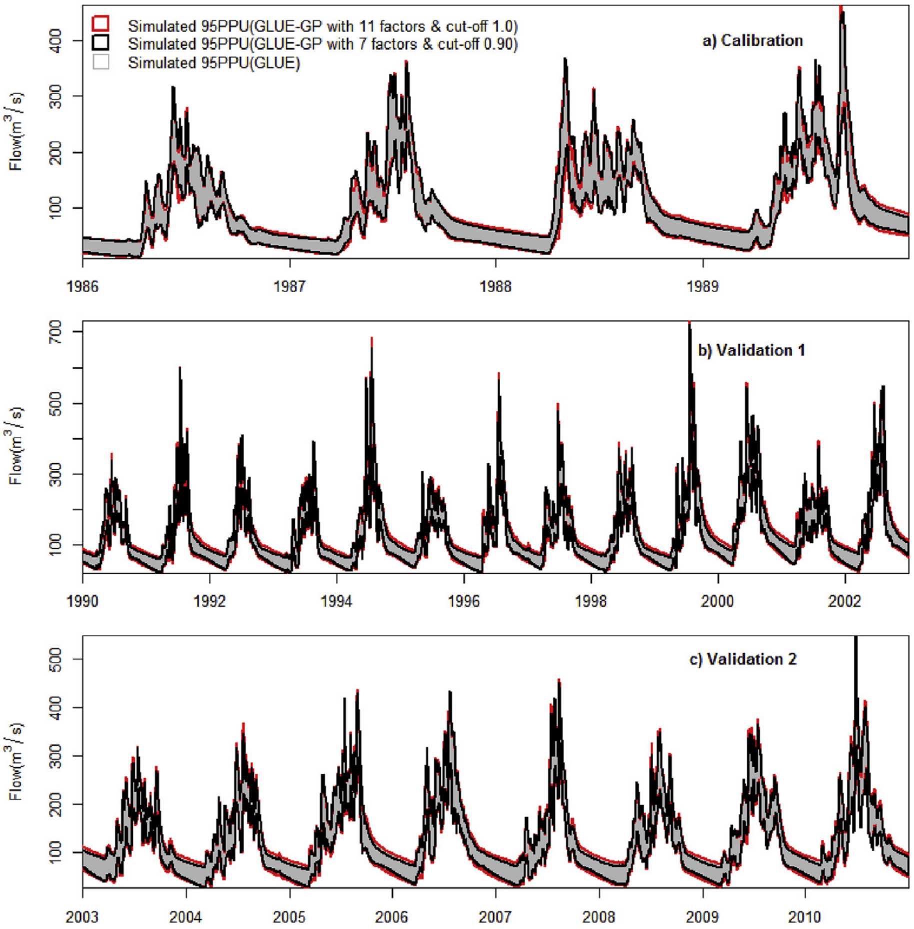

Fig. 7. Simulated 95% Prediction Uncertainty (95PPU) bands of GLUE-GP (red: 11-factor GLUE-GP with cut-off 1.0; black: 7-factor GLUE-GP with cut-off 0.90) and GLUE (grey) in the calibration (top plot) and two validation periods.

Sobol' indices (e.g., S_i, S_{Ti}, and S_{ij}) are used. These indices can rank the general importance of each factor and measure general interactions among factors. Therefore it is most useful for studying the average behavior of each factor and factor interaction, and excluding insensitive factors. The shortcoming is that it cannot provide local information of individual factors and factor interaction in factor space. The GP emulation however provides the expectations of $E(y|x_i)$ and $E(y|x_i, x_j)$, yielding insight into how different factors affect model behavior along their ranges, and their joint effects in their subspaces. This can determine how to reduce factor space in model calibration or uncertainty analysis. In this study, we examined expectations of $E(y|x_i)$ and $E(y|x_i, x_j)$ (Fig. 4). However, as first order interaction indices (S_{ij}) were rather small ($<=2\%$), we mainly used expectations of $E(y|x_i)$ to narrow factor ranges. If first order interactions (S_{ij}) are high, $E(y|x_i, x_j)$ should also be used when reducing factor ranges.

5.3. Choice of sensitive factors

The choice of sensitive factors makes some differences to the performance of GLUE-GP, as shown in Table 4 and Fig. 5. Compared with the 11-factor GLUE-GP, the 7-factor GLUE-GP has a higher sampling efficiency while possessing similar *r-factor*s and *p-factor*s, for the two higher cut-offs.

5.4. Choice of the cut-off

In the study, to illustrate the tradeoffs involved, we compared three cut-offs in the posterior expectations: 1.0, 0.95 and 0.90. When the cut-off decreases (moves towards the optimal region), it will cause a larger factor range reduction and involve more factors in the range reduction (Table 3), and then lead to increasing sampling efficiency (e.g. Table 4), in addition to a narrower uncertainty range (e.g. Fig. 6), and a more skewed histogram of the objective function towards the optimal region (e.g. Fig. 5). When the cut-off is far from the optimal region, it will not have too much effect on the factor range and will lead to a similar sampling efficiency and uncertainty result to GLUE; when it is close to the optimal region, it will reduce the factor range significantly and therefore increase sampling efficiency but narrow uncertainty. A good cut-off should be not too far from and not too close to the optimal region. An iterative approach could best start with a cut-off slightly far from the optimal region and involve several iterations in the GP construction and factor range reduction. However this will require more model runs and several GP constructions.

6. Conclusions

This paper introduces GLUE-GP, an augmented GLUE approach based on a Gaussian Process emulation, which is used to undertake a Bayesian sensitivity analysis to narrow down the factor space through reduction in the number of factors and the factor range. The application involved a semi-distributed hydrologic model SWAT in the Kaidu River Basin, using a standard root mean square error (*SRMSE*) performance, or so-called GLUE likelihood measure as demonstration. Compared to the standard GLUE, it yields significant improvements in behavioral sampling efficiency (from 3% to as much as 22% of successfully sampled behaviors) but with around half the number of samples and associated model runs, i.e., 2600 versus 5,000, while yielding not substantially different uncertainties to standard GLUE in validation mode. Consequently it locates the optimal region at a lower computational cost because the constructed GP assists in narrowing factor ranges towards the optimal factor region.

In the application of GLUE-GP, the critical step is the GP construction which involves an experimental design (i.e. efficient selection of points in the response surface of the original model), hyperparameter estimation and assessment of how well the constructed GP reproduces the original model behavior of interest. Normally, if the response surface is smooth, a small number of design points is sufficient (e.g. in this case study around 300 points are sufficient). However, as emulators aim to speed up model simulations, it should take less computational cost than the original model.

The trick of GLUE-GP is to emulate the model just well enough to allow a good sensitivity analysis, and select an appropriate cutoff in conditional expectation of the output that reduces the number of factors and their ranges to a computationally manageable level. Thus GLUE can then sample where it matters (reducing the number of samples) and obtain reasonably close uncertainty bounds to the original GLUE (which themselves are approximate in any case but hopefully a useful appreciation of predictive uncertainties). This procedure can be adapted for other uncertainty analysis techniques and models that have a smooth response surface. It should be particularly suited to models with prohibitive runtimes.

Acknowledgment

The research was supported by National Natural Science Foundation of China (41361140361), and State Key Laboratory of Desert and Oasis Ecology Project (Y471161). The authors wish to express their gratitude to manuscript editor Dr. Joseph Guillaume and four anonymous reviewers for insightful and constructive comments that helped improve the manuscript considerably.

References

Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., Srinivasan, R., 2007. Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. J. Hydrol. 333, 413–430.

Aistleitner, C., Hofer, M., Tichy, R., 2012. A central limit theorem for Latin hypercube sampling with dependence and application to exotic basket option pricing. Int. J. Theor. Appl. Finance 15 (07), 1250046.

Arnold, J.G., Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modelling. Hydrol. Process 19, 563–572. https://doi.org/10.1002/hyp.5611.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment - Part 1: model development. J. Am. Water Resour. As 34, 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Arnold, J.G., Youssef, M.A., Yen, H., White, M.J., Sheshukov, A.Y., Sadeghi, A.M., Moriasi, D.N., Steiner, J.L., Amatya, D.M., Skaggs, R.W., Haney, E.B., 2015. Hydrological processes and model representation: impact of soft data on calibration. Trans. ASABE 58 (6), 1637–1660.

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, S.L., Neitsch, S.L., 2012. Soil and Water Assessment Tool Input/Output Documentation, Version 2012. Texas Water Resources Institute, Temple, TX, USA. TR-439.

Asher, M., Croke, B.F.W., Jakeman, A.J., Peeters, L., 2015. A review of surrogate models and their application to groundwater modeling. Water Resour. Res. 51 (8), 5957–5973.

Bastos, L.S., O'Hagan, A., 2009. Diagnostics for Gaussian process emulators. Technometrics 51, 425–438.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Jakeman, A.J., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Beven, K., Binley, A., 1992. The future of distributed models -model calibration and uncertainty prediction. Hydrol. Process 6, 279–298. https://doi.org/10.1002/hyp.3360060305.

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

Beven, K., Binley, A., 2014. GLUE: 20 years on. Hydrol. Process. 28 (24), 5897–5918.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37 (1), 35–46.

Dyn, N., Levin, D., Rippa, S., 1986. Numerical procedures for surface fitting of scattered data by radial functions. SIAM J. Sci. Stat. Comput. 7, 639–659.

Emmerich, M., Giannakoglou, K.C., Naujoks, B., 2006. Single-and multiobjective evolutionary optimization assisted by gaussian random field metamodels. Evolutionary Computation. IEEE Trans. 10, 421–439.

Fang, G., Yang, J., Chen, Y., Xu, C., De Maeyer, P., 2015. Contribution of meteorological input in calibrating a distributed hydrologic model in a watershed in the Tianshan Mountains, China. Environ. Earth Sci. 74, 2413–2424. https://doi.org/10.1007/s12665-015-4244-7.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. statistics 1–67.

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

Iorgulescu, I., Beven, K.J., Musy, A., 2005. Data-based modelling of runoff and chemical tracer concentrations in the Haute–Menthue (Switzerland) research catchment. Hydrol. Process. 19, 2557–2574.

Iorgulescu, I., Beven, K.J., Musy, A., 2007. Flow, mixing, and displacement in using a data-based hydrochemical model to predict conservative tracer data. Water Resour. Res. 43, W03401 https://doi.org/10.1029/2005WR004019.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21, 602–614.

Jones, D.R., 2001. A taxonomy of global optimization methods based on response surfaces. J. Glob. Optim. 21, 345–383.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modelling: 1. Theory. Water Resour. Res. vol. 42, W03407.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B, Stat. Methodol. 425–464.

Kucherenko, S., Albrecht, D., Saltelli, A., 2016. Exploring Multi-dimensional Spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques. To be submitted for publication to Reliability Engineering & System Safety arXiv:1505.02350.

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

Liu, T., Willems, P., Pan, X.L., Bao, A.M., Chen, X., Veroustraete, F., Dong, Q.H., 2011.

Climate change impact on water resource extremes in a headwater region of the Tarim basin in China. Hydrol. Earth Syst. S. C. 15, 3511–3527. https://doi.org/10.5194/hess-15-3511-2011.

Luo, Y., Arnold, J., Allen, P., Chen, X., 2012. Baseflow simulation using SWAT model in an inland river basin in Tianshan Mountains, Northwest China. Hydrol. Earth Syst. S. C. 16, 1259–1267. https://doi.org/10.5194/hess-16-1259-2012.

Mantovan, P., Todini, E., 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. J. Hydrol. 330, 368–381.

Matott, L.S., Babendreier, J.E., Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resour. Res. 45 (6).

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

McMillan, H., Clark, M., 2009. Rainfall–runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme. Water Resour. Res. 45, W04418 https://doi.org/10.1029/2008WR007288.

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

Norton, J., 2015. An introduction to sensitivity assessment of simulation models. Environ. Model. Softw. 69, 166–174.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a bayesian approach. Journal of the royal statistical society: series B. Stat. Methodol. 66, 751–769.

Ratto, M., Young, P.C., Romanowicz, R., Pappenberger, F., Saltelli, A., Pagano, A., 2007. Uncertainty, sensitivity analysis and the role of data based mechanistic modeling in hydrology. Hydrol. Earth Syst. S. C. 11, 1249–1266.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Stat. Sci. 4, 409–423.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145, 280–297.

Saltelli, A., Tarantola, S., Chan, K.S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41, 39–56.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer. Wiley & Sons, Chichester,United Kingdom.

Setegn, S.G., Rayner, D., Melesse, A.M., Dargahi, B., Srinivasan, R., 2011. Impact of climate change on the hydroclimatology of lake tana basin, Ethiopia. Water Resour. Res. 47, W04511 https://doi.org/10.1029/2010WR009248.

Shen, Z.Y., Chen, L., Chen, T., 2012. Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: a case study of SWAT model applied to Three Gorges Reservoir Region, China. Hydrol. Earth Syst. S. C. 16, 121–132. https://doi.org/10.5194/hess-16-121-2012.

Shin, M.J., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J., 2015. A review of foundational methods for checking the structural identifiability of models: results for rainfall-runoff. J. Hydrol. 510, 1–16.

Singhee, A., Rutenbar, R.A., 2009. Novel Algorithms for Fast Statistical Analysis of Scaled Circuits, vol. 46. Springer Science & Business Media.

Sobol', I.M., 1990. On sensitivity estimation for nonlinear mathematical models. Mat. Model. 2, 112–118.

Sobol', I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. 55 (1), 271–280.

Sobol', I.M., Asotsky, D., Kreinin, A., Kucherenko, S., 2011. Construction and comparison of high-dimensional Sobol? Generators, 2011. Wilmott J. 64–79.

Song, X.M., Zhang, J.Y., Zhan, C.S., Xuan, Y.Q., Ye, M., Xu, C.G., 2015. Global sensitivity analysis in hydrological modeling: review of concepts, methods, theoretical framework, and applications. J. Hydrol. 523, 739–757. https://doi.org/10.1016/j.jhydrol.2015.02.013.

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in Peel Inlet, II, identification of critical uncertainties via generalised sensitivity analysis. Water Resour. Res. 14, 43–49.

Stedinger, J.R., Vogel, R.M., Lee, S.U., Batchelder, R., 2008. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resour. Res. 44 https://doi.org/10.1029/2008WR006822.

Todini, E., 2007. Hydrological catchment modelling: past, present and future. Hydrology Earth Syst. Sci. 11 (1), 468–482.

Vrugt, J.A., Bouten, W., 2002. Validity of first-order approximations to describe parameter uncertainty in soil hydrologic models. Soil Sci. Soc. Am. J. 66, 1740–1751.

Wiener, N., 1938. The homogeneous chaos. Am. J. Math. 60 (4), 897–936.

Xiu, D., Karniadakis, G.E., 2002. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. 24 (2), 619–644.

Yang, J., 2011. Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. Environ. Modell. Softw. 26, 444–457. https://doi.org/10.1016/j.envsoft.2010.10.007.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007. Hydrological modelling of the chaohe Basin in China: statistical model formulation and bayesian inference. J. Hydrol. 340, 167–182.

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. J. Hydrol. 358, 1–23.

Yen, H., Hoque, Y., Harmel, R.D., Jeong, J., 2015. The impact of considering uncertainty in measured calibration/validation data during auto-calibration of hydrologic and water quality models. Stoch. Environ. Res. Risk Assess. 29, 1891–1901. https://doi.org/10.1007/s00477-015-1047-z.