https://doi.org/10.5194/hess-2020-563
Preprint. Discussion started: 12 November 2020

Hydrology and Earth System Sciences
Discussions
Open Access

Quantifying input uncertainty in the calibration of water quality models: reshuffling errors via the secant method

Xia Wu[1,2], Lucy Marshall[2], Ashish Sharma[2]

[1]College of Hydrology and Water Resources, Hohai University, Nanjing, 210098, China
[2]School of Civil and Environmental Engineering, University of New South Wales, Sydney, 2052, Australia

Correspondence to: Lucy Marshall (lucy.marshall@unsw.edu.au)

Abstract. Uncertainty in inputs can significantly impair parameter estimation in water quality modeling, necessitating accurate quantification of input errors. However, decomposing input error from model residual error is still challenging. This study develops a new algorithm, referred to as Bayesian error analysis with reshuffling (BEAR), to address this problem. The basic approach requires sampling errors from a pre-estimated error distribution and then reshuffling them with their inferred ranks via the secant method. This approach is demonstrated in the case of total suspended solids (TSS) simulation via a conceptual water quality model. Based on case studies using synthetic data, the BEAR method successfully isolates the input error and parameter error. The results of a real case study demonstrate that even with the presence of model structural error and output data error, the BEAR method can approximate the true input and bring a better model fit through an effective input modification. However, its effectiveness is limited by the assumption that the input uncertainty should be dominant and that the prior information of the input error model can be estimated. The application of the BEAR method in TSS simulation is effective for understanding a range of water quality conditions and the further developed algorithm can be extended to other water quality predictions.

1 Introduction

For robust water management, uncertainty analysis is of growing importance in water quality modeling (Refsgaard et al., 2007). It can provide knowledge of error propagation and the magnitude of uncertainty impacts in model simulations to guide improved predictive performance (Radwan et al., 2004). However, the implementation of uncertainty analysis in water quality models (WQMs) is still challenging due to complex interactions among sources of multiple errors, generally caused by a simplified model structure (structural uncertainty), imperfect observed data (input uncertainty and observation uncertainty in calibration data) and limited parameter identifiability (parametric uncertainty) (Refsgaard et al., 2007).

Among them, input uncertainty is expected to be particularly significant in a WQM, interpreted here as the observation uncertainty of any input data. Observation uncertainty is different from other sources of uncertainty in modeling since these uncertainties arise independently of the WQM itself, thus, their properties (e.g. probability distribution family and distribution parameters) can, at least in principle, be estimated prior to the model calibration and simulation by analysis of the data acquisition instruments and procedures (McMillan et al., 2012). Rode and Suhr (2007) and Harmel et al. (2006) reviewed the uncertainty associated with selected water quality variables based on the empirical quality of observations. The general methodology developed in their studies can be extended to the analysis of other water quality variables. Besides the error coming from the measurement process, the error from surrogated data is another major source of input uncertainty (McMillan et al., 2012). Measurements of water quality variables often lack desirable temporal and spatial resolutions, thus, the use of surrogate or proxy data is necessary for improved inference of water quality parameters (Evans et al., 1997, Stubblefield et al., 2007). For a surrogate error, its probability distribution is easy to estimate from the residuals between the measurements and proxy values. These estimated error distributions are "prior knowledge" of input uncertainty before any model calibration and can serve as the a-priori uncertainty estimation in the modeling process.

Input uncertainty can lead to bias in parameter estimation in water quality modeling (Chaudhary and Hantush, 2017, Kleidorfer et al., 2009, Willems, 2008). Improved model calibration requires isolating the input uncertainty from the total uncertainty. However, the precise quantification of time-varying input errors is still challenging when other types of uncertainties are propagated through to the model results. In hydrological modeling, several approaches have been developed to characterize time-varying input errors, and these may hold promise for application in WQMs. The Bayesian total error analysis (BATEA) method provides a framework that has been widely used (Kavetski et al., 2006). Time-varying input errors are defined as multipliers on the input time series and inferred along with the model parameters in the Bayesian calibration scheme. It leads to a high-dimensionality problem, which restricts the application of this approach to the assumption of event-based multipliers (the same multiplier applied to one storm event). In the Integrated Bayesian Uncertainty Estimator (IBUNE) (Ajami et al., 2007) approach, multipliers are not jointly inferred with the model parameters, but sampled from the assumed distribution and then filtered by the constraints of simulation fitting. This approach reduces the dimensionality significantly and can be applied in the assumption of the data-based multiplier (one multiplier for one input data) (Ajami et al., 2007). However, this approach results in an underestimation of the multiplier variance and misidentification of the uncertainty sources (Renard et al., 2009). From the above, a new strategy should be developed to avoid high dimensional computation and ensure the accuracy of error identification.

To complete this goal, this study develops a new algorithm – Bayesian error analysis with reshuffling (BEAR). The derivation and details of the BEAR algorithm in quantifying input errors are described in Sect. 2. Section 3 introduces the build-up/wash-off model (BwMod) to illustrate this approach. Its model input, streamflow, often suffers from observational errors from a rating curve. By comparing the results with other calibration frameworks, the ability of the BEAR method is explored in a synthetic case and a real case. In this way, the new algorithm is tested in a simple situation (with an assumption of true output data and model structure) and in a realistic situation (with the interference of multiple error sources) respectively. Section 4 evaluates the BEAR method and its implementation. Finally, Section 5 outlines the main conclusions and recommendations for this work.

2 Methodology

2.1 Basic theory of identifying the input error in model calibration

A WQM in the ideal situation without any error can be described as

$$\boldsymbol{Y}^* = M(\boldsymbol{X}^* \mid \boldsymbol{\theta}^*) \tag{1}$$

where the true output Y^* is simulated by the perfect model M with the true inputs $\boldsymbol{X}^*$ and the true model parameters $\boldsymbol{\theta}^*$. Here and in the following contents, a capital bold letter (e.g. $\boldsymbol{X}, \boldsymbol{Y}$) represents a vector and a lower case (e.g. x, y) represents a variable.

In reality, the model input $\boldsymbol{X}^o$ (typically the rainfall or streamflow in a WQM) inevitably suffers from input error $\boldsymbol{\varepsilon}_X$. This will result in a calibrated model parameter $\boldsymbol{\theta}^c$ biased from the true value $\boldsymbol{\theta}^*$ (Kleidorfer et al., 2009). Thus, under the assumption that the input errors are additive to the true input data $\boldsymbol{X}^*$ and the output data and model structure are generally without errors, the model residual $\boldsymbol{\varepsilon}$ in a traditional calibration can be described by

$$\boldsymbol{\varepsilon} = \boldsymbol{Y}^o - \boldsymbol{Y}^s = \boldsymbol{Y}^* - M(\boldsymbol{X}^* + \boldsymbol{\varepsilon}_X \mid \boldsymbol{\theta}^c) \tag{2}$$

Under the ideal situation without input errors, the residual will reduce to zero, like

$$\boldsymbol{\varepsilon} = \boldsymbol{Y}^o - \boldsymbol{Y}^s = \boldsymbol{Y}^* - M(\boldsymbol{X}^* \mid \boldsymbol{\theta}^*) = 0 \tag{3}$$

To counter the influence of input errors in a traditional calibration, an appealing approach is to subtract estimated errors $\boldsymbol{\varepsilon}_X^p$ from the observed input $\boldsymbol{X}^o$. This is illustrated as the "proposed" approach and the superscript p represents the values in this "proposed" approach. The residual $\boldsymbol{\varepsilon}^p$ will change to

$$\boldsymbol{\varepsilon}^P = \boldsymbol{Y}^o - \boldsymbol{Y}^p = \boldsymbol{Y}^* - M(\boldsymbol{X}^p \mid \boldsymbol{\theta}^P) = \boldsymbol{Y}^* - M(\boldsymbol{X}^* + \boldsymbol{\varepsilon}_X - \boldsymbol{\varepsilon}_X^p \mid \boldsymbol{\theta}^P) \tag{4}$$

If the equivalence between $\boldsymbol{\varepsilon}_X$ and $\boldsymbol{\varepsilon}_X^p$ can be ensured for each data point, the modified input $\boldsymbol{X}^p$ then becomes the same as the true value $\boldsymbol{X}^*$. The proposed calibration (Eq. (4)) will result in an ideal calibration (Eq. (1)), where the optimal parameters $\boldsymbol{\theta}^p$ will converge to their true values $\boldsymbol{\theta}^*$ and the model residual $\boldsymbol{\varepsilon}^P$ will decrease to zero. Thus, the precise identification of input errors will result in the ideal model parameters and minimized residual error.

Selecting the optimal input error series according to the statistical characteristics of the residual error is the basic theory of current methods (Ajami et al., 2007, Kavetski et al., 2006). The challenge is to optimize the input errors effectively when parameter errors impact this optimization. The BATEA method considers input errors and model parameters as a whole and infers them by taking advantage of the correlation among them. This results in a high dimensionality problem that cannot be avoided (Renard et al., 2009). The IBUNE method makes use of the stochasticity of sampled errors and selects the most suitable error and parameter sample to minimize the residual error. This is less effective because the probability of co-occurrence of all optimal error/parameter values is very low (Renard et al., 2009). An improved strategy is necessary to properly infer input errors through minimising total model residual error.

2.2 The innovation of the secant method

The secant method can be applied to address this problem. This is an iterative process to produce better approximations to the roots of a real-valued equation (Ralston and Jennrich, 1978). Here, the root is the optimal value of each input error and the equation is the corresponding model residual equal to zero. A traditional approach to updating this is impractical because the estimated input error will fully complement the model error and always lead to a zero residual error regardless of the model parameters. More discussion on this is stated in Sect. 4.1.

This study attempts to transform this optimization into the rank domain. Here, the rank is defined as the order of any individual value relative to the other sampled values, and determines the relative magnitude of each error in all data errors. For most WQM studies, the probability distribution of any input errors $f(\eta_\varepsilon)$ can be estimated as per prior information. If there is knowledge of the error distribution, the error value only depends on its rank in this distribution and this error distribution can then constrain the value range of sampled errors. Therefore, the secant method is very useful in the rank domain, where the root turns to the optimal rank of each input error (rather than its value) and the equation is still the corresponding model residual equal to zero. This new approach, referred to as the Bayesian error analysis with reshuffling (BEAR) method, should be implemented in two steps: sampling the errors from the estimated error distribution and reshuffling these sampled errors corresponding to the inferred error ranks via the secant method.

The secant method (Ralston and Jennrich, 1978) can be repeated as

$$k_{i,q} = k_{i,q-1} - \varepsilon_{i,q-1}^{p} \frac{k_{i,q-1} - k_{i,q-2}}{\varepsilon_{i,q-1}^{p} - \varepsilon_{i,q-2}^{p}} \tag{5}$$

until a sufficiently accurate target value is reached. In this study, the target value is a residual of zero ($\varepsilon_{i,q}^{p} = 0$) indicating a perfect model fit (with input errors estimated exactly). Here, $k_{i,q}$ represents the estimated rank for ith input error at the qth iteration, $\varepsilon_{i,q-1}^{p}$ is the residuals corresponding to the input error rank $k_{i,q-1}$. The error rank of each data point is updated respectively via Eq. (5), where i =1,…n. n is the data length and also the number of the estimated errors as these errors are data-based.

After calculating Eq. (5), it is possible that the rank $k_{i,q}$ is out of the rank range (for example, less than 1 or more than n), or not an integer. Sorting $k_{i,q}$ in all the ranks $k_{i,q}(i=1,...,n)$ can address this problem by effectively scaling the calculated ranks

$k_{i,q}$ to an integer from 1 to n. Thus, in Eq. (5), $k_{i,q}$ should be changed to $K_{i,q}$, representing the pre-rank. After sorting $K_{i,q}$ for all the errors, the post-rank $k_{i,q}$ will then belong to reasonable values.

From the above, estimating the rank of input errors via the secant method can be described as the following two steps:

Update the rank of each input error $K_{i,q} \quad (i=1,\dots,n)$ via the secant method respectively:

$$K_{i,q} = k_{i,q-1} - \varepsilon_{i,q-1}^{p} \frac{k_{i,q-1} - k_{i,q-2}}{\varepsilon_{i,q-1}^{p} - \varepsilon_{i,q-2}^{p}} \tag{6}$$

Sorting $K_{i,q} \quad (i=1,\dots,n)$ in all the error pre-ranks $\boldsymbol{K_q}$ to obtain a reasonable rank:

$$k_{i,q} = k(K_{i,q}) \tag{7}$$

where k( ) means calculating its rank.

2.3 Approximate Bayesian Computation - Sequential Monte Carlo (ABC – SMC)

This study chooses Approximate Bayesian Computation via Sequential Monte Carlo (ABC–SMC) as the calibration scheme. ABC-SMC was first proposed by Sisson et al. (2007) and developed in the research of Toni et al. (2008). The ABC method is especially useful for problems in which the likelihood function is analytically intractable or costly to compute in traditional Bayesian approaches. For formal Bayesian approaches,, the likelihood function must be set carefully to meet the assumption about the residual error distribution, and this setting impacts the parameter estimation (Smith et al., 2015, McInerney et al., 2017, Wu et al., 2019). In the ABC method, setting an objective function is more general allowing for potentially complex input error distributions where the likelihood is difficult to write.

In the ABC–SMC approach, the parameter $\boldsymbol{\theta}^p$ is first sampled from a prior distribution $P\left(\boldsymbol{\theta}^p\right)$ (referred to as population 1). Then it is propagated through a sequence of intermediate distributions $P(\boldsymbol{\theta}^p \left| OF(\boldsymbol{Y}^o, \boldsymbol{Y}^p) \le \tau_s \right.)$, s=1,…, F-1 (referred to as intermediate population 2, …, F-1), until it represents a sample from the target distribution $P(\boldsymbol{\theta}^p \left| OF(\boldsymbol{Y}^o, \boldsymbol{Y}^p) \le \tau_F \right.)$ (referred to as the posterior distribution). The tolerance τ_s of the objective function is chosen that τ_1>…>τ_F >0, thus the distributions sequentially evolve towards the target posterior.

2.4 Algorithm and an example of the BEAR method

According to the previous derivations, the algorithm quantifying input errors via the BEAR method is demonstrated in Fig. 1 and an illustrative example is presented in Table 1 and Fig. 2. Based on an ABC-SMC calibration scheme, the BEAR method works by replacing the observed input with a modified input that is obtained through the estimated input error rank via the secant method. In Fig. 1, s refers to the number of the sequential updating populations in the ABC-SMC scheme, which increases until the objective function (measuring the fit between the calibration data and model outputs) of the sth population is less than the final tolerance τ. The final tolerance τ (i.e. the stopping criterion) is difficult to set before calibration due to the unknown range of objective function values, but in practice, it can be estimated after several population calibrations, according to the actual calculation range of the objective function and the target accuracy. In this study, the calibration stops when 1000 proposed parameter sets are rejected in a row. The first tolerance τ_1 should be set sufficiently large to start the update. Any intermediate tolerance τ_s is set as the 30% quantile of the objective function results of the previous population s-1, such that it reduces automatically with a new population calculation.

In each calibration population, the input error ranks are updated over q iterations, where q increases until the objective function is less than tolerance τ_s. When q=1 and q=2, the input errors are randomly sampled from the estimated error distribution because two sets of samples are prerequisites for the updating via the secant method (Table 1). Regarding these, a series of error ranks $\boldsymbol{k}_q^p$, modified inputs $\boldsymbol{X}_q^p$, model outputs $\boldsymbol{Y}_q^p$, and model residuals $\boldsymbol{\varepsilon}_q^p$ are calculated, demonstrated as the 1st and 2nd iteration in Table 1. In later iterations (q>=3), the error rank $\boldsymbol{k}_q^p$ is updated via the secant method (Eq. (6) and (7)), demonstrated in the first two columns in the 3rd and 4th iteration in Table 1. According to the new rank $k_{i,q}^p$, the value with the same rank in the 2nd iteration is the estimated error in the new iteration. For example, the new rank at the 1st time step in the 3rd iteration is 6, and its corresponding value in the 2nd iteration is -0.02, therefore, -0.02 is set as the updated input error at the 1st time step in the 3rd iteration. After the same reshuffling strategy, the re-ranked input errors will then lead to a new series of the modified inputs $\boldsymbol{X}_q^p$, model outputs $\boldsymbol{Y}_q^p$ and model residuals $\boldsymbol{\varepsilon}_q^p$.

Note however if the model parameters are far away from the true values, especially in the initial population, iterative updating of the error ranks will have little effect in reducing the model residual. Therefore, the maximum times of iterations should be set, referred to as Q. If q exceeds Q, the algorithm returns to the step resampling the model parameters (seen in Fig. 1). An example of four iterations is demonstrated in Table 1 and Fig. 2.

In the example given in Table 1, before reshuffling errors (i.e. the 1st iteration and 2nd iteration), the input errors do not approach the true values shown in Fig. 2, having much larger objective function results than the 3rd and 4th iteration. After the error reshuffling, the objective function calculated in the 4th iteration is smaller than the result in the 3rd iteration, illustrating that the estimated errors in the 4th iteration are closer to the true values than the 3rd iteration. This is also supported by Fig. 2 where the red line (4th iteration) has a stronger correlation with the black line (true input error) than the yellow line (3rd iteration). From the above, the true input errors can be approximated through updating the error ranks to minimize the objective function of the residuals.

2.5 Comparison with other methods

To evaluate the ability of the BEAR method in quantifying input errors, three methods are compared, denoted as method T, D, R. Method "T" is the "traditional" method, regarding the observed input as error-free without identifying input errors (i.e. Eq. (2)), while the other two methods employ a latent variable to counteract the impacts of input error and build the modified input (i.e. Eq. (4)). In method D, "D" refers to the probability "Distribution" of input error, which is additional information considered in the calibration. This error distribution can be estimated before calibration according to the studies in the introduction. Especially in the context of proxy errors, the probability distribution can be easily calculated via the residuals between the measurements and the corresponding proxy values. From this error distribution, potential input errors are randomly sampled and filtered by the minimization of the objective function, which is similar to the basic framework of the IBUNE method (Ajami et al., 2007). Method R represents the BEAR method developed in this study. "R" refers to the "Reshuffling" strategy via the secant method, which is an additional process to that used in method D to improve the input error quantification.

3 Case studies

3.1 Water quality model: the build-up/wash-off model (BwMod)

This study tests the BEAR algorithm in the context of the build-up/wash-off model (BwMod), which is a group of models to simulate two processes in sediment dynamics, including the build-up of sediments during dry periods and the wash-off process during wet periods. The two formulations were developed in a small-scale experiment (Sartor and Boyd, 1972), while in applications at the catchment scale, the conceptualized parameters largely abandon their physical meanings and the formulations can be considered a "black-box" (Bonhomme and Petrucci, 2017). This study chooses Eq. (8) to describe the build-up process and Eq. (9) to express the wash-off of sediments, representing the non-linear relationship between the wash-off load (output) and the runoff-rate (input). These two equations were applied in the research of Sikorska et al. (2015) and in this study, are written in the MATLAB programming language with the integration of the BEAR method. The time scale is typically set as daily, and the spatial scale is set as the catchment in this study. This version of BwMod has four parameters (Table 2). The model input is streamflow, which typically comes from the observation of a rating curve. As discussed in the introduction, the error distribution can be estimated prior to the model calibration via a rating curve analysis. The output of the BwMod is the concentration of total suspended solids (TSS), whose transport can be efficiently simulated by the conceptualization of the build-up/wash-off process (Bonhomme and Petrucci, 2017, Sikorska et al., 2015). Although BwMod is relatively simple compared with process-based WQMs, its nonlinearity and the use of surrogates for the input data can make it a typical WQM scenario to test the BEAR algorithm.
The overall BwMod equations are:

$$\frac{dS_{a,t}}{dt} = \kappa \cdot \left(S_{max} - S_{a,t}\right) - s\left(S_{a,t}\right)\text{肜} \tag{8}$$

where $S_{a,t}$ (kg) is the sediment amount available on the catchment surface to be washed-off at time t; $s\left(S_{a,t}\right)$ (kg· s^{-1}) is the amount of sediment in the stream at time t, described by the function

$$s(S_{a,t}) = a \cdot Q_t^b \cdot S_{a,t} \tag{9}$$

The output TSS concentration $C_{TSS,t}$ (kg · m^{-3}) is derived via:

$$C_{TSS,t} = \frac{s\left(S_{a,t}\right)}{Q_t} \tag{10}$$

3.2 Case study 1: Synthetic data

First, the BEAR method is tested in a controlled situation with synthetic data, where the model is affected only by input errors and parameter errors. The true input $\boldsymbol{X}^*$ is set as the daily streamflow data of the catchment in the real case (USGS ID: 04087030), covering 1095 days from 2009/10/01 to 2012/09/29. The observed input $\boldsymbol{X}^o$ is generated based on two types of input error models: an additive formulation and a multiplicative formulation, and the errors are assumed to follow a normal distribution with mean μ as 0.2 and standard deviation (SD) σ as 0.5. An additive formulation (denoted as 'add' in Table 3) is suitable to illustrate error generation, while the multiplicative formulation (denoted as 'mul' in Table 3) is specifically applied for errors induced from a log-log regression procedure, which is common for water quality proxy processes (Rode and Suhr, 2007). In the additive formulation, the generated input may be negative. If so, the negative input should be truncated to a positive value. In the multiplicative formulation, the generated input will stay positive. The true output $\boldsymbol{Y}^*$ is the simulated TSS concentration via BwMod corresponding to the true input $\boldsymbol{X}^*$ and model parameters set as the reference values in

. The observed output $\boldsymbol{Y}^o$ is assumed to be the same as the true simulation $\boldsymbol{Y}^*$, i.e. without error.

In the calibration, the objective function is set as the Mean Squared Error (MSE). Considering the unknown initial sediment loads in real applications, the calibration sets 90 days as a warm-up period to remove the influence of antecedent conditions. Following the algorithm described in Sect. 2.4, the model parameters and the time-varying input errors are estimated. In each population of the ABC-SMC calibration scheme, 50 sets of model parameters are updated. In the first population, the model parameters are sampled from a uniform distribution with the prior range described in Table 2.

The prior information about error parameters (i.e. σ and μ) contains two conditions: one is fixed as the reference values (denoted as 'fixed' in Table 3), the other one is given the prior range, which needs to infer the error parameters in the calibration (denoted as '*inferred*' in Table 3).

To sum up, this study considers four scenarios in the synthetic case, including two sets of synthetic data generating from two input error models and two types of prior information about the error parameter (the details are shown in Table 3). Each scenario is calibrated via method T, method D and method R respectively. Their algorithms are described in Sect. 2.5 and their results are compared in Fig. 3 and Fig. 4. Figure 3 shows the statistical characteristics of the overall estimations. Figure 4 demonstrates the temporal dynamics of input estimations and model simulations.

Evaluating the input error quantification, method R always has much higher correlations with the true error series than method D in all calibration scenarios (shown in Fig. 3(3)). When the error parameters are inferred, the estimations of σ via method D are smaller than the reference value (shown in Fig. 3(1)). This conclusion has also been reported in the study of Renard et al. (2009). The reason for this is that the randomness of the likelihood function leads to an underestimation of the SD of input errors. Compared with method D, the σ estimation via method R is less biased from its true value (shown in Fig. 3 (1)), while the estimation of μ is worse via method R (shown in Fig. 3(2)).

For the model simulation, method R always produces the best output fit in all scenarios, supported by the highest red boxplots in Fig. 3(4). Also in Fig. 4, regardless of the calibration scenarios, the output uncertainty bands of method R (red parts) almost overlaps the true output (green line), being much better than method T (pink parts) and method D (blue parts). However, the input uncertainty bands vary depending on the calibration scenarios. When the error parameters are fixed at the reference values (in the scenarios *add-fixed* and *mul-fixed*), method R always outperforms the other two methods regardless of input error models, as its Nash–Sutcliffe efficiency coefficient (NSE) are the highest (shown in Fig. 3(5)). In Fig. 4(1) and Fig. 4(3), the input uncertainty bands of method R (red parts) generally converge to the true value (green line), being better than method D (blue parts). Without the reshuffling strategy, Method D even gives worse input estimation and model simulations than method T, demonstrated by the lower blue boxplots than pink boxplots in Fig. 3(5)) and Fig. 3(4). This illustrates that the ill-posed error sources in method D exert a negative impact on the model simulations. When the error parameters are inferred (in the scenarios of *add-inferred* and *mul-inferred*), the performance of method R depends on the input error models. For the scenario of *add-inferred*, method R is still better than other methods, having the biggest NSE (shown in Fig. 3(5)) and the closest error parameter estimation to the reference value (shown in Fig. 3(1) and Fig. 3(2)), although the input uncertainty band is more negatively biased from the true value (green line) than method D in Fig. 4(2). For the scenario of *mul-inferred*, the modified inputs via method R are further from the reference value than method D (shown in Fig. 3(5)), which might result from worse μ estimations for the input error (shown in Fig. 3(2)).

https://doi.org/10.5194/hess-2020-563
Preprint. Discussion started: 12 November 2020

3.3 Case study 2: Real data

The above synthetic case only exhibits input error and parameter error, which focuses on testing the ability of the BEAR method in quantifying time-varying input errors while estimating model parameters. In real-life applications, the impacts of model structural error and output data error cannot be ignored. In order to explore the BEAR method in more general situations, e.g. with other errors' interference, a real case of one catchment located in southeast Wisconsin, USA is demonstrated. Table 4 is a description of the test catchment and data (Baldwin et al., 2013). The daily TSS concentration and streamflow data are collected from the USGS database on National Real-Time Water Quality (https://nrtwq.usgs.gov/).

The daily streamflow data in the USGS database comes from a stage-streamflow rating curve, where the stage and streamflow form a log-log linear relationship and the streamflow proxy errors follow a normal distribution with μ as 0 and σ as 0.103. This prior information is used in the real calibration, denoted as *O-fixed* scenario in Table 3. Because the BEAR method is implemented under the assumption that the input uncertainty is so significant that other sources of uncertainties can be ignored, another input data source with more significant data uncertainty, the streamflow simulation from a hydrological model, has been considered. This study selects GR4J (Perrin et al., 2003) as the hydrological model and calibrates its parameters with the USGS streamflow data as calibration data. If the USGS streamflow data is regarded as the true input data, the residual error after the model calibration can approximate the data error of GR4J simulation, which follows a normal distribution in log space with μ as 0 and σ as 0.764. The BwMod calibration using this input data source and the prior information on data error is denoted as *S-fixed* scenario in Table 3. To explore the ability of the BEAR method in other situations where the prior information about the input error is not sufficient, two scenarios with a wider range of the error parameters has also been considered, denoted as *O-inferred* and *S-inferred* in Table 3. The real case is also calibrated via three methods (i.e. method T, method D and method R) and adopts the same setting of the calibration algorithm as the synthetic case.

Figure 5 uses several statistics to evaluate the calibration scenarios. For all scenarios in Fig. 5(b1), method R always produces a better fit to the output data than method D, consistent with the synthetic case shown in Fig. 3(4). In Fig. 5(b2), "Reliability" here is the ratio of observations caught by the confidence interval of 5%-95%, and the average width of this interval band is referred to as "Sharpness" (Yadav et al., 2007, Smith et al., 2010). In the *S-fixed* and S-inferred scenarios with significant input errors, the results of method R show much higher reliability with a larger sharpness. However, in the *O-fixed* scenario with insignificant input errors (i.e. σ =0.103), the reliability and sharpness of method R are smaller than method D. Fig. 5(a) demonstrates that the σ estimations vary depending on the calibration methods, but stay almost identical between two data sources. This illustrates that the impacts of other sources of errors significantly impair the error quantification, and their impacts are varied for different methods.

In the real case shown in Fig. 6, method R still produces the best fit to the output and the uncertainty band of the modified input via method D is centered on the observed data. In Fig. 6(c), the uncertainty bands of the modified input are consistent in all scenarios except the *O-fixed* scenario with insignificant input errors (i.e. σ =0.103). The uncertainty bands are closer to the observed streamflow (green line), even in (c3) and (c4) where the input data comes from the simulated streamflow (black line). According to the results of method T in Fig. 6(a), the simulations corresponding to the observed streamflow (in (a1) and (a2)) catch the dynamics of observed TSS concentration better than the simulations corresponding to the simulated streamflow (in (a3) and (a4)). Here, the observed streamflow from the rating curve should be closer to the true input data, and could be regarded as the reference value. Given that, the modified inputs via method R are more reasonable.

4 Discussion

4.1 The effectiveness of rank estimation

The novelty of the BEAR method lies in transforming a direct error value estimation to an error rank estimation. In a continuous sequence of data, the potential error values have an infinite number of combinations, while the error rank has limited combinations, dependent on the data length. It is far more efficient to estimate the error rank than estimate the error value. Compared with the IBUNE framework (Ajami et al., 2007), the BEAR method additionally infers the error ranks to adjust the order of the sampled errors and reduce their randomness, which significantly improves the accuracy of the error estimation (as demonstrated by much higher NSEs than method D in Fig. 3). The application of the secant method plays an essential role in this by inferring each error rank according to the residual error.

Note that modifying each input error according to the corresponding residual error only works in the rank domain. In the value domain, if there is no constraint on the estimated input errors, they will fully compensate for the residual error with the aim of minimizing the objective function and subsequently be overfitted. There are two ways to impose restrictions. One is to regard errors and model parameters as a whole in calibration, resulting in the high dimensional computation (Kavetski et al., 2006). The other is to sample error randomly from the assumed error model IBUNE (Ajami et al., 2007), whose precision cannot be guaranteed. While in the rank domain, the value range of the sampled errors can be effectively limited by the assumed error model.

One thing to note in the rank estimation is that even corresponding to the same rank, the error sampled at different times could be largely different, especially for a small sample size (depending on the data length) or a large standard deviation of the assumed error distribution. This problem can be addressed by selecting the optimal solution from multiple samples according to the minimization of an appropriate objective function. The secant method is a successive approximation algorithm and one single iteration cannot guarantee the optimal results. Considering these two points, the BEAR method set q iterations in the algorithm (Fig. 1). q increasing until the objective function becomes smaller than the tolerance.

4.2 The impacts of prior information of input error model

Method D employs the same framework as IBUNE (Ajami et al., 2007), taking advantage of stochastic error samples to modify the input observations. In Fig. 4 and Fig. 6, the uncertainty bands of modified inputs (blue parts) encompass the input observations (black line), illustrating that the intrinsic quality of the input observation determines the algorithm performance.

Figure 6 demonstrates that if the input error is insignificant in the residual, like in the *O-fixed* and *O-inferred* scenarios of the real case, the resultant simulations will fit the observed output (green line) well. Otherwise, the simulations are far away from the observed outputs (black line) due to inaccurate input observations (in the *S-fixed* and *S-inferred* scenarios in the real case). As per the finding in the previous study of Renard et al. (2010), if the SD of input errors is inferred with the model parameters, method D will underestimate the SD (Fig. 3(1) and Fig. 5(a2)). If the intrinsic SD of input errors is large, a fixed SD cannot improve the input modification and model simulation, demonstrated by a wider band in Fig. 6(b3) than in Fig. 6(b4). If the SD of input errors is small, the prior information will constrain the impacts of other sources of errors. From the above, the data quality is more important than the availability of prior information for method D, especially when the intrinsic SD of the input error is large.

However, the findings in method R are quite different. Although method R infers the input error by minimizing the model residual error, it is much more effective than method D to minimise the residual errors. For the synthetic case (Fig. 4(c)) and real case (Fig. 6(c)), the model simulations via method R (red parts) are very close to the output observations (green line). In other words, the estimated input error mainly depends on the output observations. Therefore, in the real case with the same output observation (Fig. 6(c)), the modified inputs are consistent among the scenarios. Given this, the model structure error plays an important role in estimating the input error.

To constrain the impacts of the other sources of error, accurate prior information about the input error model is important in method R. In the synthetic case, fixed scenarios always produce a higher NSE of the modified input (Fig. 3(5)) and a larger correlation in the estimation error (Fig. 3(5)) than inferred scenarios. This illustrates that prior information can limit the impacts of model parameter error. In Fig. 6(a1), the modified inputs in the real case are around the reference value (green line), while in Fig. 6(a2), the modified inputs are biased from the reference value (green line). It should be noted that this difference is obvious in the scenarios with insignificant input error (where the model structural error is relatively large). When the input error is dominant, like the *S-inferred* scenario, method R becomes more effective to estimate the input error, bringing a more precise estimation of the error SD than the *O-inferred* scenario and similar results to the *S-fixed* scenario.

To sum up, for method R, an accurate input error model can constrain the adverse impacts of the other sources of errors, especially when the other sources of error are dominant. But for method D, the input data quality is more important than this prior information.

4.3 The extension to other modeling scenarios

In this study, the BEAR method was developed in the calibration of BwMod at the daily time scale, whose input and output correspond at each time step. Therefore, in Eq. (6), the model residual $\varepsilon_{i,q-1}^{p}$ and input error rank $k_{i,q-1}$ are at the same time step i. If the water quality system exhibits response delays, the time lag between the forcing data and the response (described as the lag) should be considered in the algorithm and Eq. (6) needs to be modified as per Eq. (11).

$$K_{i,q} = k_{i,q-1} - \varepsilon^{p}_{i+lag,q-1} \frac{k_{i,q-1} - k_{i,q-2}}{\varepsilon^{p}_{i+lag,q-1} - \varepsilon^{p}_{i+lag,q-2}} \tag{11}$$

If the response caused by an input is not instantaneous but exhibits persistence (i.e. occurs over several time steps), the autocorrelation in the output should be addressed to ensure the independence assumption of the rank updating is satisfied. Current ways to deal with this problem in hydrologic modeling can provide a reference in the potential modification of the BEAR method. The persistence of residual errors can be represented by an autoregressive moving average (ARMA) model (Kuczera, 1983) or autoregressive (AR) model (Schaefli et al., 2007, Bates and Campbell, 2001). However, the ability of these approaches needs further discussion in systems with correlated responses.

5 Conclusion

Taking advantage of the prior information of an input error model, a new method, Bayesian error analysis with reshuffling (BEAR), is developed to approach the time-varying input errors in WQM inference. Through the investigation of synthetic data and real data, this method is shown to be robust and effective. The novelties of this algorithm are: (1) Estimating the error rank rather than directly estimating the error value which significantly improves the effectiveness of input error quantification by reducing the potential search space for input errors. (2) The modification of the secant method links the error rank of each input data to its corresponding residual, which addresses the high dimensionality problem in current calibration methods.

However, the work in this study still identifies a few areas needing to be explored. Firstly, the availability of prior knowledge of the input error model is important. When this information is not reliable or even cannot be estimated, a significant issue is the selection of a suitable error assumption. Thus, a general measure should be found to judge whether an error model is appropriate, especially in real cases where the "true" information is limited. Secondly, extensions of the BEAR method to other water quality modeling scenarios are subject to problems such as delayed and autocorrelated responses. Related studies in hydrologic modeling to deal with the delay and persistency of responses could be references in the modification of the BEAR method. Thirdly, if the sampling and reshuffling strategy is developed within a more comprehensive framework to quantify multiple sources of error, the interactions amongst these error sources might be well identified and the quantification of individual errors might be improved. This study provides a starting point for developing the rank estimation via the secant method to identify input error. Further study is necessary to modify the algorithm and improve confidence in extended case studies or model scenarios.

Code/Data availability

The daily streamflow and TSS concentration data for real case catchment (ID: USGS 04087030) can be accessed by the National Real-Time Water Quality website of USGS, the link is https://nrtwq.usgs.gov/.

Author contribution

Lucy Marshall and Ashish Sharma designed the research. Xia Wu developed the research code, analyzed the results, and prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Australian Research Council [FT120100269] and the Australian Research Council (ARC) Discovery Award [DP170103959] to Dr. Marshall.

References

AJAMI, N. K., DUAN, Q. & SOROOSHIAN, S. 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water resources research*, 43.

BALDWIN, A. K., ROBERTSON, D. M., SAAD, D. A. & MAGRUDER, C. 2013. Refinement of Regression Models to Estimate Real-Time Concentrations of Contaminants in the Menomonee River Drainage Basin, Southeast Wisconsin, 2008–11. *US Geological Survey Scientific Investigations Report 2013-5174.* US Geological Survey Reston, Virginia.

BATES, B. C. & CAMPBELL, E. P. 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water resources research*, 37**,** 937-947.

CHAUDHARY, A. & HANTUSH, M. M. 2017. Bayesian Monte Carlo and maximum likelihood approach for uncertainty estimation and risk management: Application to lake oxygen recovery model. *Water Research*, 108**,** 301-311.

EVANS, J., WASS, P. & HODGSON, P. 1997. Integrated continuous water quality monitoring for the LOIS river syndromme. *Science of the total environment,* 194**,** 111-118.

HARMEL, R., COOPER, R., SLADE, R., HANEY, R. & ARNOLD, J. 2006. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Transactions of the ASABE,* 49**,** 689-701.

KAVETSKI, D., KUCZERA, G. & FRANKS, S. W. 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water resources research,* 42.

KLEIDORFER, M., DELETIC, A., FLETCHER, T. & RAUCH, W. 2009. Impact of input data uncertainties on urban stormwater model parameters. *Water Science and Technology,* 60**,** 1545-1554.

KUCZERA, G. 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. *Water Resources Research,* 19**,** 1151-1162.

MCINERNEY, D., THYER, M., KAVETSKI, D., LERAT, J. & KUCZERA, G. 2017. Improving probabilistic prediction of daily streamflow by identifying P areto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research,* 53**,** 2199-2239.

MCMILLAN, H., KRUEGER, T. & FREER, J. 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes,* 26**,** 4078-4111.

PERRIN, C., MICHEL, C. & ANDRÉASSIAN, V. 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology,* 279**,** 275-289.

RADWAN, M., WILLEMS, P. & BERLAMONT, J. 2004. Sensitivity and uncertainty analysis for river quality modelling. *Journal of Hydroinformatics,* 6**,** 83-99.

RALSTON, M. L. & JENNRICH, R. I. 1978. Dud, A Derivative-Free Algorithm for Nonlinear Least Squares. *Technometrics,* 20**,** 7-14.

REFSGAARD, J. C., VAN DER SLUIJS, J. P., HØJBERG, A. L. & VANROLLEGHEM, P. A. 2007. Uncertainty in the environmental modelling process – A framework and guidance. *Environmental Modelling & Software,* 22**,** 1543-1556.

RENARD, B., KAVETSKI, D. & KUCZERA, G. 2009. Comment on “An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction” by Newsha K. Ajami et al. *Water Resources Research,* 45.

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.

RODE, M. & SUHR, U. 2007. Uncertainties in selected river water quality data.

SCHAEFLI, B., TALAMBA, D. B. & MUSY, A. 2007. Quantifying hydrological modeling errors through a mixture of normal distributions. *Journal of Hydrology,* 332**,** 303-315.

SISSON, S. A., FAN, Y. & TANAKA, M. M. 2007. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences,* 104**,** 1760-1765.

SMITH, T., MARSHALL, L. & SHARMA, A. 2015. Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology,* 528**,** 29-37.

SMITH, T., SHARMA, A., MARSHALL, L., MEHROTRA, R. & SISSON, S. 2010. Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resources Research,* 46.

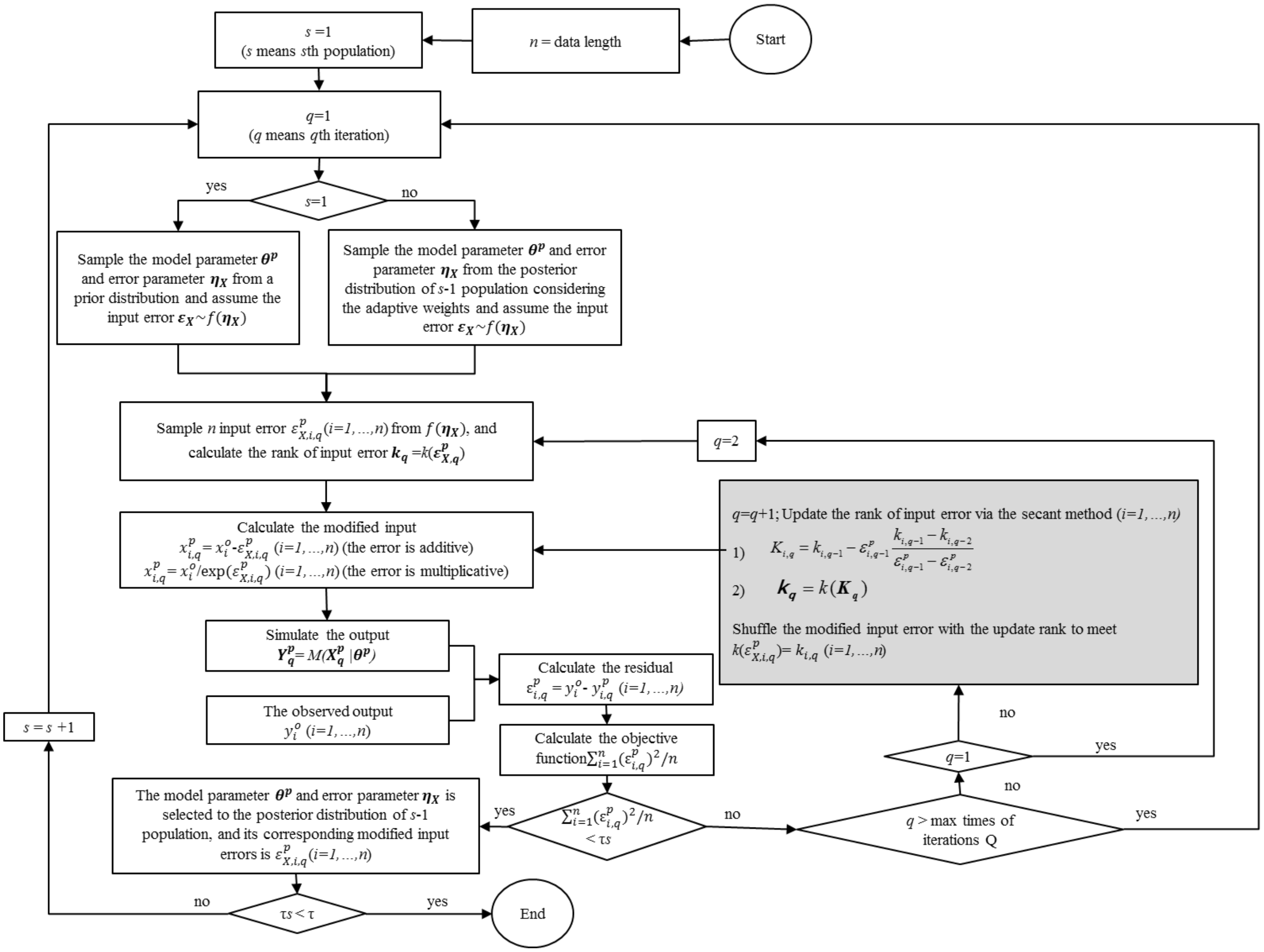
STUBBLEFIELD, A. P., REUTER, J. E., DAHLGREN, R. A. & GOLDMAN, C. R. 2007. Use of turbidometry to characterize suspended sediment and phosphorus fluxes in the Lake Tahoe basin, California, USA. *Hydrological Processes,* 21**,** 281-291.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. & STUMPF, M. P. 2008. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface,* 6**,** 187-202.

WILLEMS, P. 2008. Quantification and relative comparison of different types of uncertainties in sewer water quality modeling. *Water Research,* 42**,** 3539-3551.

WU, X., MARSHALL, L. & SHARMA, A. 2019. The influence of data transformations in simulating Total Suspended Solids using Bayesian inference. *Environmental Modelling & Software,* 121**,** 104493.

YADAV, M., WAGENER, T. & GUPTA, H. 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources,* 30**,** 1756-1774.


Start
n = data length
s =1 (s means sth population)
q=1 (q means qth iteration)
s=1
yes
no
Sample the model parameter $\boldsymbol{\theta^p}$ and error parameter $\boldsymbol{\eta_X}$ from a prior distribution and assume the input error $\boldsymbol{\varepsilon_X} \sim f(\boldsymbol{\eta_X})$
Sample the model parameter $\boldsymbol{\theta^p}$ and error parameter $\boldsymbol{\eta_X}$ from the posterior distribution of s-1 population considering the adaptive weights and assume the input error $\boldsymbol{\varepsilon_X} \sim f(\boldsymbol{\eta_X})$
Sample n input error $\varepsilon^p_{X,i,q}$(i=1, ...,n) from $f(\boldsymbol{\eta_X})$, and calculate the rank of input error $\boldsymbol{k_q}$ =k($\boldsymbol{\varepsilon^p_{X,q}}$)
q=2
Calculate the modified input
$x^p_{i,q}$ = x^o_i-$\varepsilon^p_{X,i,q}$ (i=1, ...,n) (the error is additive)
$x^p_{i,q}$ = x^o_i/exp($\varepsilon^p_{X,i,q}$) (i=1, ...,n) (the error is multiplicative)
q=q+1; Update the rank of input error via the secant method (i=1, ...,n)
1) $K_{i,q} = k_{i,q-1} - \varepsilon^p_{i,q-1} \frac{k_{i,q-1} - k_{i,q-2}}{\varepsilon^p_{i,q-1} - \varepsilon^p_{i,q-2}}$
2) $\boldsymbol{k_q} = k(\boldsymbol{K_q})$
Shuffle the modified input error with the update rank to meet $k(\varepsilon^p_{X,i,q})$= $k_{i,q}$ (i=1, ...,n)
Simulate the output $\boldsymbol{Y^p_q}$= M($\boldsymbol{X^p_q}$ |$\boldsymbol{\theta^p}$)
The observed output y^o_i (i=1, ...,n)
Calculate the residual $\varepsilon^p_{i,q}$ = y^o_i- $y^p_{i,q}$ (i=1, ...,n)
Calculate the objective function$\sum^n_{i=1}(\varepsilon^p_{i,q})^2/n$
no
q=1
yes
s = s +1
The model parameter $\boldsymbol{\theta^p}$ and error parameter $\boldsymbol{\eta_X}$ is selected to the posterior distribution of s-1 population, and its corresponding modified input errors is $\varepsilon^p_{X,i,q}$(i=1, ...,n)
yes
$\sum^n_{i=1}(\varepsilon^p_{i,q})^2/n$ < τs
no
no
q > max times of iterations Q
yes
no
τs < τ
yes
End

450

Figure 1: Flowchart of the algorithm to quantify the input errors via Bayesian error analysis with reshuffling (BEAR) method

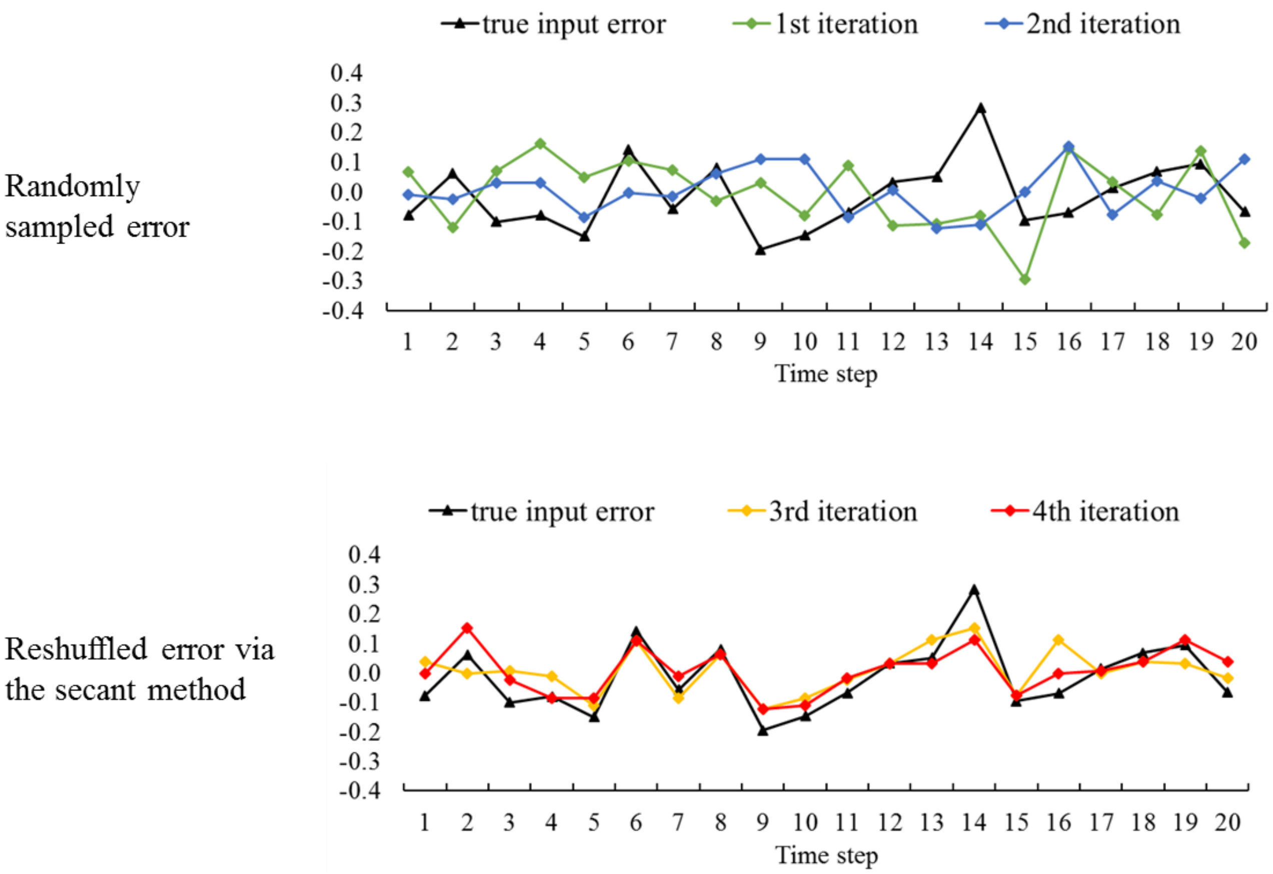


Figure 2: Demonstration of the results in Table 1 before and after reshuffling the errors via the secant method

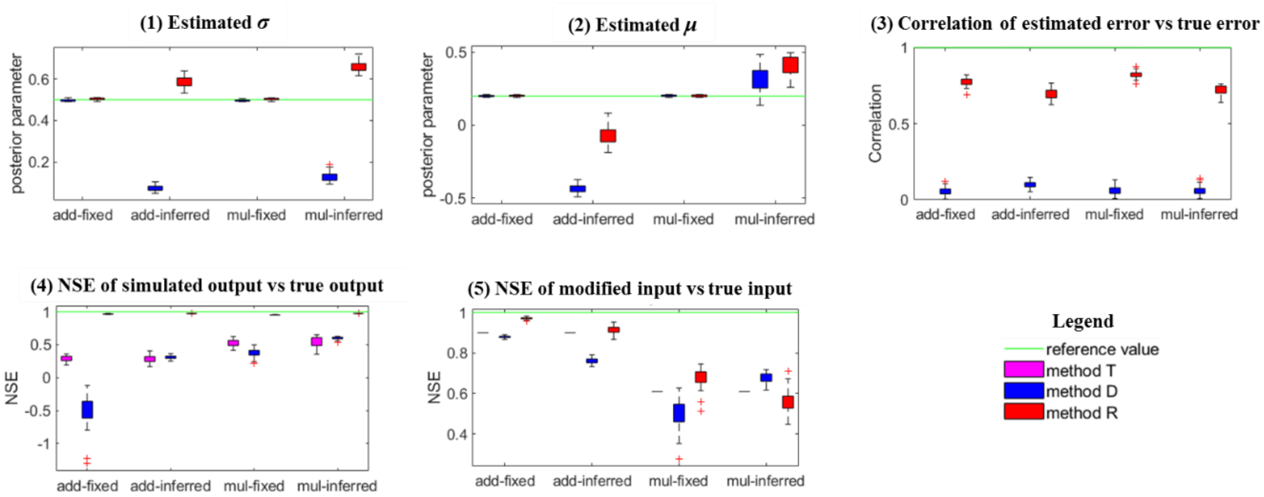


Figure 3: Comparison of statistical characteristics of four calibration scenarios in the synthetic case (including *add-fixed*, *add-inferred*, *mul-fixed* and *mul-inferred*; notations are given in Table 3) via three calibration methods (including method T, method D and method R, their algorithms are explained in Sect. 2.5)

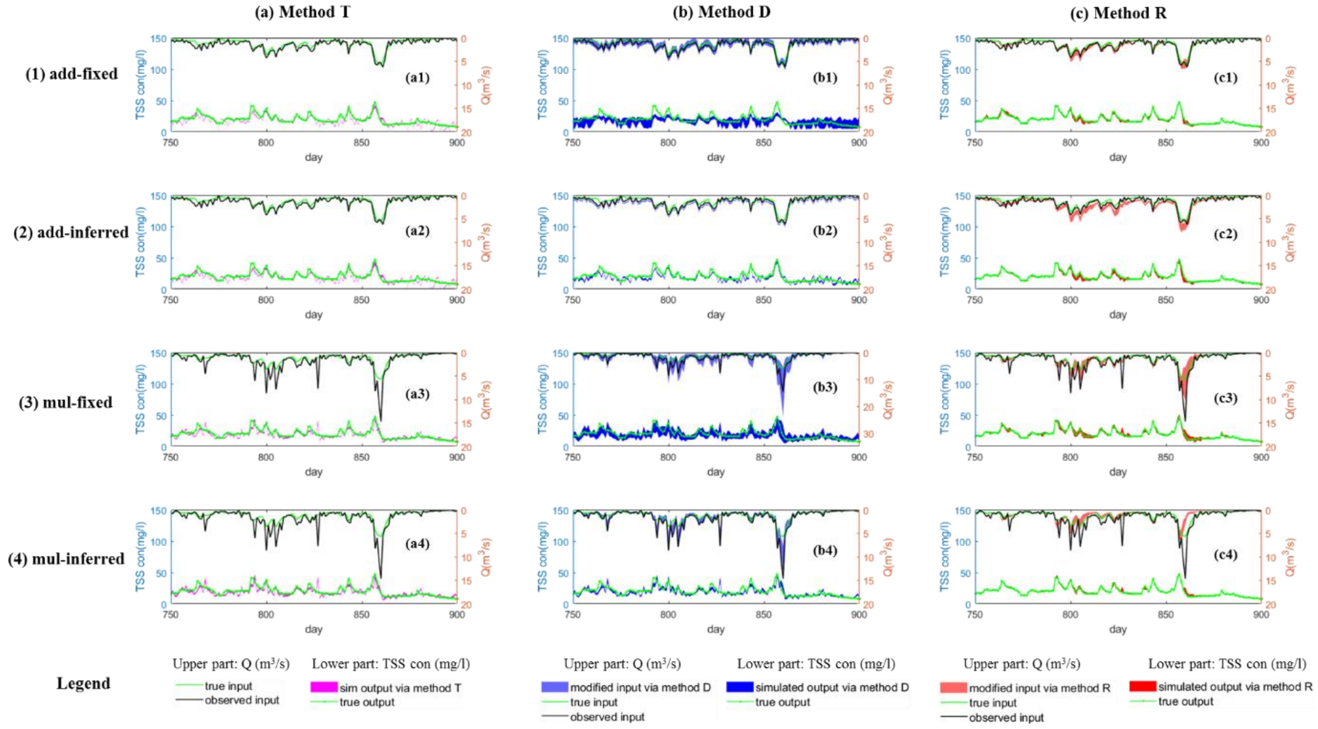


Figure 4:Comparison of time series of synthetic data and uncertainty bands estimated via three calibration methods (including method T, method D and method R; algorithms are explained in Sect. 2.5) for a select period of four calibration scenarios in the synthetic case (including *add-fixed*, *add-inferred*, *mul-fixed* and *mul-inferred*; notations are given in Table 3)

https://doi.org/10.5194/hess-2020-563
Preprint. Discussion started: 12 November 2020

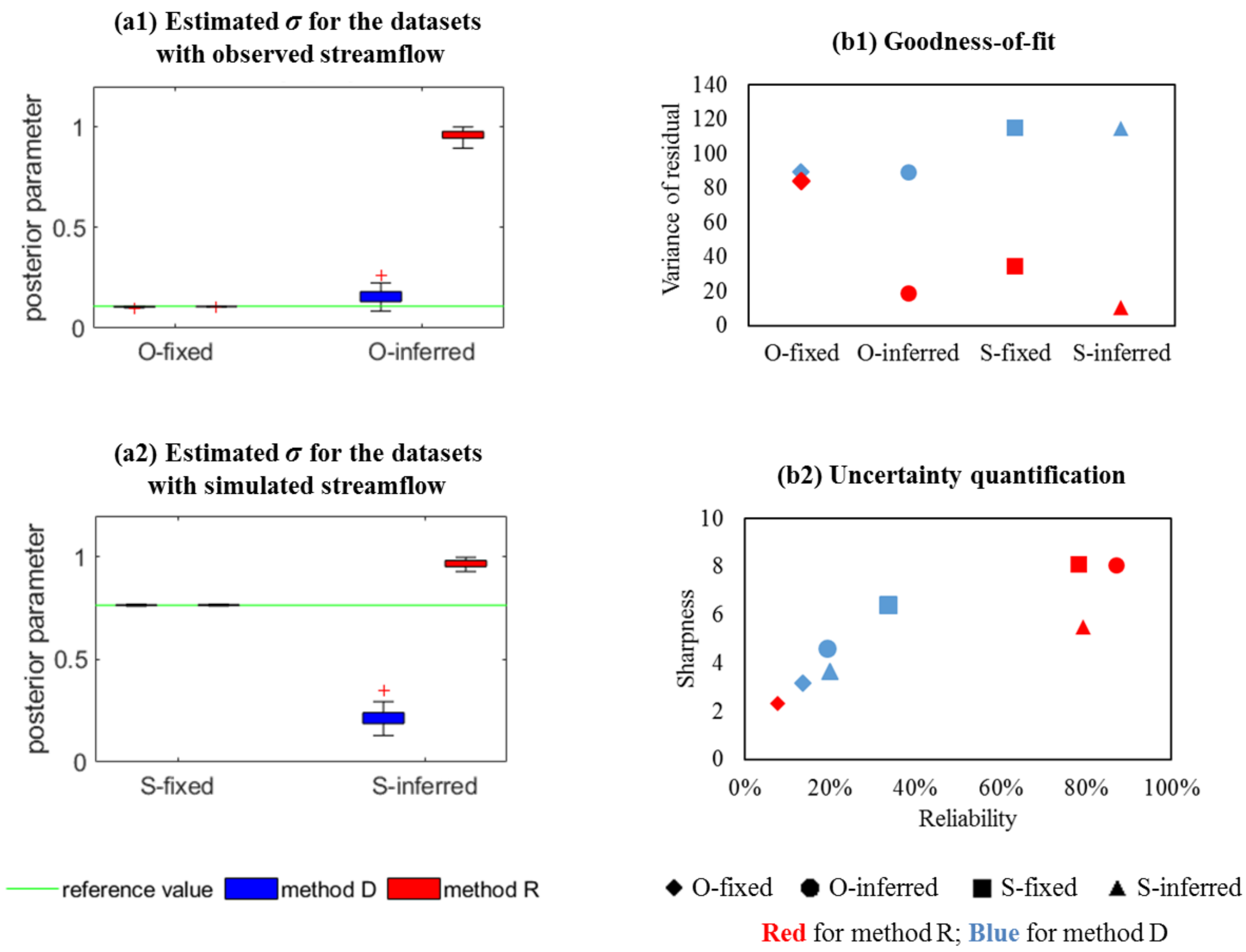


Figure 5:Comparison of statistical characteristics of four calibration scenarios in the real case (including *O-fixed*, *O-inferred*, *S-fixed* and *S-inferred*, their notations are given in Table 3) via three calibration methods (including method T, method D and method R, their algorithms are explained in Sect. 2.5)

470

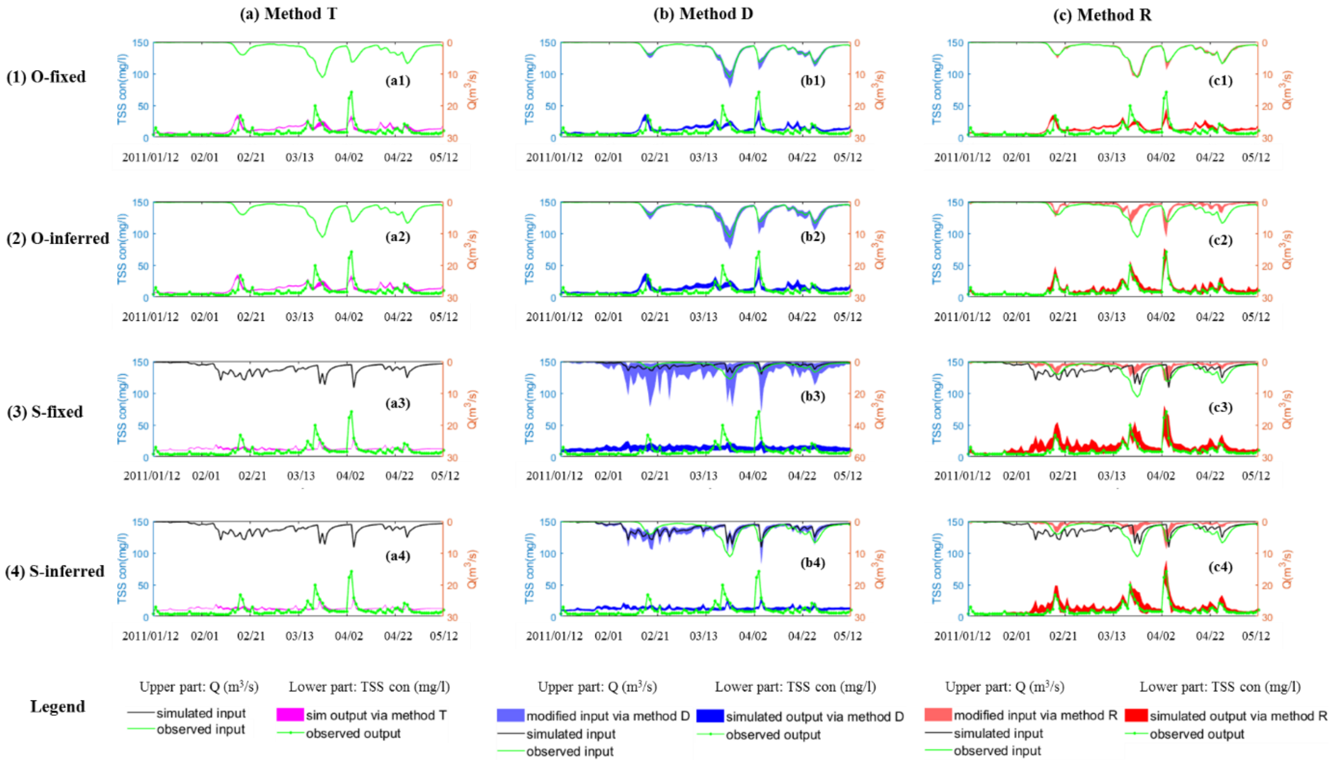


Figure 6:Comparison of time series of real data and uncertainty bands estimated via three calibration methods (including method T, method D and method R, algorithms are explained in Sect. 2.5) for a select period of four calibration scenarios in the real case (including *O-fixed*, *O-inferred*, *S-fixed* and *S-inferred*, notations are given in Table 3)

Table 1-1 An example illustrating the rank updating approach via the secant method

Time Step *i*	Observed data		1st iteration (random sample)					2nd iteration (random sample)				
	x_i^o	y_i^*	$\varepsilon_{X,i,1}^p$	$k_{i,1}$	$x_{i,1}^p$	$y_{i,1}^p$	$\varepsilon_{i,1}^p$	$\varepsilon_{X,i,2}^p$	$k_{i,2}$	$x_{i,2}^p$	$y_{i,2}^p$	$\varepsilon_{i,2}^p$
1	2.24	24.1	0.07	13	2.18	23.8	0.29	-0.01	9	2.25	24.0	0.13
2	1.87	23.6	-0.12	3	1.99	24.0	-0.49	-0.02	6	1.90	23.8	-0.23
3	1.37	23.1	0.07	14	1.30	22.5	0.58	0.03	14	1.34	22.6	0.43
4	1.02	22.2	0.16	20	0.86	21.2	0.98	0.03	13	0.99	21.7	0.41
5	0.90	22.2	0.05	12	0.85	21.4	0.78	-0.09	3	0.98	22.0	0.21
6	0.99	21.5	0.10	17	0.89	21.8	-0.29	0.00	10	0.99	22.2	-0.70
7	0.76	21.5	0.07	15	0.69	20.8	0.66	-0.02	8	0.78	21.2	0.23
8	0.87	21.4	-0.03	9	0.90	22.0	-0.59	0.06	16	0.81	21.5	-0.09
9	0.60	21.4	0.03	10	0.57	20.1	1.31	0.11	17	0.49	19.5	1.88
10	0.62	21.3	-0.08	7	0.70	21.0	0.31	0.11	18	0.51	19.8	1.52
11	0.70	21.3	0.09	16	0.61	20.4	0.87	-0.09	4	0.78	21.5	-0.20
12	0.85	21.6	-0.11	4	0.97	22.4	-0.76	0.01	12	0.85	21.8	-0.17
13	1.55	24.2	-0.11	5	1.66	24.7	-0.46	-0.12	1	1.67	24.7	-0.53
14	3.20	27.2	-0.08	6	3.28	27.7	-0.54	-0.11	2	3.31	27.8	-0.60
15	1.91	24.6	-0.29	1	2.21	24.9	-0.25	0.00	11	1.91	24.2	0.43
16	1.51	23.6	0.14	19	1.37	22.8	0.80	0.15	20	1.36	22.9	0.72
17	1.26	22.7	0.03	11	1.23	22.7	0.07	-0.08	5	1.34	23.1	-0.36
18	1.09	22.1	-0.08	8	1.16	22.6	-0.56	0.04	15	1.05	22.2	-0.12
19	1.06	22.0	0.14	18	0.92	21.8	0.23	-0.02	7	1.08	22.5	-0.47
20	0.98	22.4	-0.17	2	1.15	22.8	-0.40	0.11	19	0.87	21.6	0.82
		Objective function $\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_{i,q})^2$					0.40					0.47

https://doi.org/10.5194/hess-2020-563
Preprint. Discussion started: 12 November 2020

Table 1-2 An example illustrating the rank updating approach via the secant method

Time Step *i*	3rd iteration (the secant method)						4th iteration (the secant method)					
	$K_{i,3}$	$k_{i,3}$	$\varepsilon^p_{X,i,3}$	$x^p_{i,3}$	$y^p_{i,3}$	$\varepsilon^p_{i,3}$	$K_{i,4}$	$k_{i,4}$	$\varepsilon^p_{X,i,4}$	$x^p_{i,4}$	$y^p_{i,4}$	$\varepsilon^p_{i,4}$
1	5.63	6	-0.02	2.21	23.9	0.23	13.17	11	0.00	2.24	24.0	0.15
2	8.76	10	0.00	1.88	23.8	-0.20	34.76	20	0.15	1.72	23.3	-0.20
3	14.00	12	0.01	1.36	22.7	0.34	4.65	7	-0.02	1.39	22.9	0.19
4	8.08	9	-0.01	1.03	21.9	0.24	3.26	4	-0.09	1.11	22.2	-0.07
5	-0.33	2	-0.11	1.01	22.1	0.12	0.72	3	-0.09	0.98	22.0	0.24
6	21.84	17	0.11	0.88	21.7	-0.19	19.51	17	0.11	0.88	21.7	-0.18
7	4.28	4	-0.09	0.85	21.6	-0.14	5.54	9	-0.01	0.77	21.2	0.24
8	17.25	16	0.06	0.81	21.5	-0.08	16.00	16	0.06	0.81	21.5	-0.10
9	-6.12	1	-0.12	0.72	21.0	0.40	-3.32	1	-0.12	0.72	21.0	0.39
10	4.18	3	-0.09	0.70	21.0	0.31	-0.87	2	-0.11	0.73	21.1	0.17
11	6.29	7	-0.02	0.72	21.1	0.22	5.44	8	-0.02	0.71	21.0	0.26
12	14.38	14	0.03	0.82	21.6	-0.03	14.36	13	0.03	0.82	21.6	-0.03
13	30.82	19	0.11	1.44	24.0	0.17	14.54	14	0.03	1.52	24.3	-0.07
14	41.98	20	0.15	3.05	27.5	-0.26	33.77	19	0.11	3.09	27.5	-0.30
15	4.71	5	-0.08	1.99	24.6	0.09	3.46	5	-0.08	1.99	24.5	0.12
16	29.64	18	0.11	1.40	23.1	0.55	11.63	10	0.00	1.52	23.4	0.22
17	10.06	11	0.00	1.26	22.8	-0.11	13.56	12	0.01	1.25	22.8	-0.03
18	16.83	15	0.04	1.05	22.2	-0.14	15.00	15	0.04	1.05	22.2	-0.13
19	14.37	13	0.03	1.02	22.2	-0.27	20.79	18	0.11	0.95	21.9	0.08
20	7.60	8	-0.02	1.00	22.2	0.23	3.80	6	0.04	0.94	22.0	0.44
Objective function $\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_{i,q})^2$						0.06						0.04

Note: $x^p_{i,q} = x^o_{i,q} - \varepsilon^p_{X,i,q}$, $y^p_{i,q} = M(x^p_{i,q} \mid \theta^p)$, M is BwMod with the model parameter θ^p (a=0.04, b=1.6,$\kappa = 0.1$, *Smax*=70000), $\varepsilon^p_{i,q} = y^*_i - y^p_{i,q}$.

In 1st and 2nd iteration: $\varepsilon^p_{X,i,1}$ and $\varepsilon^p_{X,i,2}$ are randomly sampled from N(0,0.01), $k_{i,q} = k(\varepsilon^p_{X,i,q})$.

In 3rd and latter iterations: $K_{i,q} = k_{i,q-1} - \varepsilon^p_{i,q-1}\dfrac{k_{i,q-1} - k_{i,q-2}}{\varepsilon^p_{i,q-1} - \varepsilon^p_{i,q-2}}$; $k_{i,q} = k(K_{i,q})$; $\varepsilon^p_{X,i,q}$ is $\varepsilon^p_{X,j,2}$ shuffled with $k_{i,q}$ to meet

$k_{i,q} = k(\varepsilon^p_{X,j,2}) = k(\varepsilon^p_{X,i,q})$

https://doi.org/10.5194/hess-2020-563
Preprint. Discussion started: 12 November 2020

Table 2 Descriptions of BwMod parameters

Model	Parameter	Description	Unit	Reference value	Prior range
BwMod	*a*	wash-off coefficient	-	0.04	(0, 2)
	b	wash-off exponent	-	1.6	(0, 3)
	κ	sediment accumulate rate	-	0.1	(0, 1)
	Smax	maximum amount of sediment possible to be accumulated	kg	7000	(0, 15000)

Table 3 Summary of the calibration scenarios in case studies

Scenario in the synthetic case	Notation	Input error model in the synthetic data generation	Prior information of input error model in calibration
1	*add-fixed*	$\boldsymbol{X}^o = \boldsymbol{X}^* + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0.2, 0.5^2)$	$\boldsymbol{X}^o = \boldsymbol{X}^* + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0.2, 0.5^2)$
2	*add-inferred*		$\boldsymbol{X}^o = \boldsymbol{X}^* + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mu, \sigma^2), \mu \in (-0.5, 0.5), \sigma \in (0,5)$
3	*mul-fixed*	$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0.2, 0.5^2)$	$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0.2, 0.5^2)$
4	*mul-inferred*		$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(\mu, \sigma^2), \mu \in (-0.5, 0.5), \sigma \in (0,5)$
Scenario in the real case	Notation	Input data source in the real case	Prior information of input error model in calibration
1	*O-fixed*	Observations from the rating curve (USGS database)	$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0, \sigma^2), \sigma \in (0.10, 0.11)$
2	*O-inferred*		$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0, \sigma^2), \sigma \in (0,1)$
3	*S-fixed*	Simulations from a hydrological model	$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0, \sigma^2), \sigma \in (0.76, 0.77)$
4	*S-inferred*		$\boldsymbol{X}^o = \boldsymbol{X}^* \exp(\boldsymbol{\varepsilon}), \boldsymbol{\varepsilon} \sim N(0, \sigma^2), \sigma \in (0,1)$

Table 4 Characteristics of the study catchments and calibration data

USGS station number	location	State	Drainage area (km^2)
04087030	Menomonee River at Menomonee Fall	Wisconsin, USA	89.83

land use			Period of Data	Number of Data (days)
Urban (percent)	Agricultural (percent)	Natural (percent)		
35	38	27	2009/10/01 - 2012/09/29	1095