Advances in Water Resources 88 (2016) 109–123

Contents lists available at ScienceDirect

Advances in Water Resources

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

Multiple-response Bayesian calibration of watershed water quality models with significant input and model structure errors

Feng Han[a,b], Yi Zheng[a,b,*]

[a] College of Engineering, Peking University, Beijing 100871, PR China
[b] Institute of Water Sciences, Peking University, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 18 March 2015
Revised 15 October 2015
Accepted 3 December 2015
Available online 11 December 2015

Keywords:
Water quality modeling
Non-point source pollution
Bayesian inference
Markov chain Monte Carlo
Multiple-response calibration
Uncertainty analysis

ABSTRACT

While watershed water quality (WWQ) models have been widely used to support water quality management, their profound modeling uncertainty remains an unaddressed issue. Data assimilation via Bayesian calibration is a promising solution to the uncertainty, but has been rarely practiced for WWQ modeling. This study applied multiple-response Bayesian calibration (MRBC) to SWAT, a classic WWQ model, using the nitrate pollution in the Newport Bay Watershed (southern California, USA) as the study case. How typical input and model structure errors would impact modeling uncertainty, parameter identification and management decision-making was systematically investigated through both synthetic and real-situation modeling cases. The main study findings include: (1) with an efficient sampling scheme, MRBC is applicable to WWQ modeling in characterizing its parametric and predictive uncertainties; (2) incorporating hydrology responses, which are less susceptible to input and model structure errors than water quality responses, can improve the Bayesian calibration results and benefit potential modeling-based management decisions; and (3) the value of MRBC to modeling-based decision-making essentially depends on pollution severity, management objective and decision maker's risk tolerance.

1. Introduction

Watershed water quality (WWQ) models, such as Soil and Water Assessment Tool (SWAT) [1,2], Watershed Analysis Risk Management Framework (WARMF) [3–5], and Hydrological Simulation Program - Fortran (HSPF) [6], can provide spatially and temporally distributed simulations of hydrology and water quality variables. WWQ modeling helps enhance our understanding of watershed processes of pollutants, and build explicit linkages between pollution causes and water quality effects. The models have been widely used in addressing water quality management issues [7–11], such as Total Maximum Daily Loads (TMDLs) planning [5]. However, their applications suffered significant modeling uncertainty resulting from inaccurate forcing inputs, model structural inadequacy, uncertain model parameters and observational errors (e.g., errors in water quality measurements) [12–16]. During the past decade, modeling uncertainty has been extensively discussed with regard to hydrology [17], but has received much less attention for water quality [16,18].

Bayesian inference using Markov chain Monte Carlo (MCMC) sampling is an advanced approach for model calibration and uncertainty analysis, which requires an explicit statistical model of residuals (i.e., error model) for rigorous likelihood evaluation. It provides posterior parameter distributions and can be used to assess both parametric and predictive uncertainties. In the field of hydrological modeling, various Bayesian inference approaches have been proposed and applied [19–26], but their applications to WWQ modeling have been limited [27–29]. WWQ modeling integrates pollution simulation with hydrological simulation. However, in existing WWQ models, watershed processes of soil erosion, chemical reactions and pollutant transport are often accounted for by simple equations, which reflects the current knowledge gaps [12–14]. Thus, WWQ modeling generally involves much higher model structure errors than pure hydrological modeling. On the other hand, in WWQ modeling, model inputs of both point sources (e.g., wastewater discharge) and non-point sources (e.g., fertilizer application and atmospheric deposition) loadings are often highly inaccurate due to the lack of loading data at desired temporal and spatial resolutions. For example, it is impractical to continuously monitor chemical concentrations in effluents, and in many cases only yearly or monthly loading estimates are available for daily-step simulations. Another real-world example is that bookkeeping of fertilizer uses is usually poor, and therefore it is impossible to know exactly when and where the historical fertilizer application occurred. The amount and timing of the application are often estimated based on sales data and/or plant growth cycle.

* Corresponding author at: Room 1001, Wangkezhen Building, Peking University, Beijing 100871, China. Tel./fax: +86 10 62768255.
E-mail address: yizheng@pku.edu.cn (Y. Zheng).

http://dx.doi.org/10.1016/j.advwatres.2015.12.007
0309-1708/

Therefore, WWQ modeling also involves much higher input errors than pure hydrological modeling.

Water quality observations (e.g., instream nitrate concentrations) are critical to the calibration of WWQ models, but are often scarce and poorly measured. In contrast, high-frequency (e.g., daily) flow observations at gaging stations are more available and reliable, representing a general supplement to water quality observations. However, using multiple types of observational data to constrain water quality simulation has been a highly ad hoc practice. A recent study [30] conducted a Bayesian calibration for sediment modeling using both flow and sediment observations. It implemented several sequential calibration (i.e., calibrate the model for the flow response first, and then for the sediment response) strategies in parallel, and then fused the results through an ensemble approach. This sequential calibration approach is sound and useful, but a more straightforward alternative would be multiple-response Bayesian calibration. A few studies have shown that including multiple hydrological responses (e.g., flow, soil moisture, etc.) in the Bayesian calibration can improve identifiability of model parameters and adequateness of uncertainty assessment [31,32]. As discussed before, water quality simulation in general involves greater and more complicated modeling uncertainty than hydrological simulation. Thus, the successful experience of multiple-response Bayesian calibration in hydrological modeling may not be fully transferable to WWQ modeling, and further studies are highly desired.

In this study, we investigated the impact of input and model structure errors on model calibration, prediction and management decisions. We considered a particular modeling situation, in which input and model structure errors are very significant and would severely bias the simulation results. This situation is very common in watershed-scale modeling of poorly monitored water quality parameters (e.g., nutrients, metals, pesticides, etc.), but has been rarely investigated in a multiple-response Bayesian calibration context. SWAT and Differential Evolution Adaptive Metropolis ($DREAM_{(ZS)}$), a state-of-the-art MCMC algorithm developed in the field of hydrological modeling [25,33,34], were employed as the WWQ model and MCMC algorithm, respectively. The nitrate pollution in Newport Bay watershed (Southern California, USA) was the study case. A series of numerical experiments were designed and implemented. Overall, the study demonstrated the critical role of input and model structure errors in assessing modeling uncertainty, identifying posterior parameter distributions and making management decisions, and demonstrated the feasibility and importance of performing multiple-response calibration for WWQ models in a management context. It has also been concluded that interpretation of the modeling uncertainty would depend on water quality management concerns.

2. Multiple-response Bayesian calibration

The relationship between a model output (i.e., model response) and its corresponding observation can be expressed as

$$\mathbf{Z} = \mathbf{Y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Z}$ and $\mathbf{Y}(\boldsymbol{\theta})$ are the observed and simulated values of the concerned response; $\boldsymbol{\theta}$ is a vector of uncertain model parameters; and $\boldsymbol{\varepsilon}$ is a lumped residual error term. Bayes' rule can be adopted, as

$$p(\boldsymbol{\theta}|\mathbf{Z}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}, \mathbf{Z}) \tag{2}$$

where $p(\boldsymbol{\theta}|\mathbf{Z})$ and $p(\boldsymbol{\theta})$ represent the posterior and prior distributions of $\boldsymbol{\theta}$, respectively; and $L(\boldsymbol{\theta}, \mathbf{Z})$ represents the likelihood function mathematically determined by the error model of $\boldsymbol{\varepsilon}$. In the context of multiple-response calibration, different error models are required for different responses, and therefore multiple likelihood functions are needed [31]. If the residual errors of multiple responses are assumed to be independent, the combined likelihood function (denoted as $L_{multiple}$) is the product of individual likelihood functions [21,23,31]:

$$L_{multiple} = \prod_{i=1} L^i\left(\boldsymbol{\theta}, \mathbf{Z}^i\right) \tag{3}$$

where i indicates the ith response, and $\boldsymbol{\theta}$ is a common set of random model parameters for the multiple responses.

To derive the individual and combined likelihood functions, it is critical to determine the error model of each response. In flow modeling, normal error models have often been assumed for $\boldsymbol{\varepsilon}$ (e.g., [20,24,34]), but have been criticized for being unrealistic [35]. Recent studies proposed more realistic but complex error models [19,22,26,36–38] for flow modeling. For example, in an auto-correlated, heteroscedastic, and non-Gaussian error model [22], the heteroscedasticity is reflected by a linear equation, $\sigma_t = \sigma_0 + \sigma_1 y_t$, where y_t is the simulated flow at time t, σ_t is the estimated standard deviation of the residual error ε_t, and σ_0 and σ_1 are two unknown hyper-parameters. Auto-correlation of residuals is depicted by a first order autoregressive model (i.e., AR(1)), and another hyper-parameter, the lag-1 autoregressive parameter ϕ, is then included. A skew exponential power (SEP) distribution was employed to deal with the skewness and heavy tail issue. The SEP distribution contains two tunable parameters, the skewness parameter ξ and the kurtosis parameter β. More details about the SEP distribution can be found by Schoups and Vrugt [22]. Hence, there are five hyper-parameters (σ_0, σ_1, ϕ, ξ, β), hereafter denoted as a vector $\boldsymbol{\varphi}$, to be inferred in the Bayesian calibration. This error model leads to the following log-likelihood function [22,37]:

$$log\, L\left(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z}\right) = n log \frac{2\sigma_\xi \omega_\beta}{\xi + \xi^{-1}} - \sum_{t=1}^{n} \log \sigma_t - c_\beta \sum_{t=1}^{n} |a_{\xi,t}|^{2/(1+\beta)} \tag{4}$$

where n is the number of observations; $a_{\xi,t} = \xi^{-sign(\mu_\xi + \sigma_\xi a_t)}(\mu_\xi + \sigma_\xi a_t)$, with a_t being an independent and identically distributed random error with zero mean and unit standard deviation; and μ_ξ, σ_ξ, ω_β and c_β are all functions of the skewness parameter ξ and the kurtosis parameter β (see [22] for details). According to Evin et al. [37], applying the AR(1) model to standardized residuals (i.e., $\eta_t = \frac{\varepsilon_t}{\sigma_t}$) instead of raw heteroscedastic residuals can lead to more stable predictive distributions. We followed Evin et al.'s approach and a_t is calculated as

$$a_t = \frac{\eta_t - \phi\eta_{t-1}}{\sqrt{1-\phi^2}} \tag{5}$$

A few studies have specifically discussed error models for water quality responses in the context of Bayesian inference. By Wellen et al. [30], normality and independence of error were simply assumed for sediment modeling. As suggested by Schoups and Vrugt [22], in Bayesian calibration, one could gradually increase complexity of the error model, from a simple Gaussian, homoscedastic and non-autocorrelated model to Eq. (4), until posterior checks confirmed that the residual errors are consistent with the error model assumptions.

3. Data and methods

3.1. Newport Bay Watershed

The Newport Bay Watershed (NBW) is located in Orange County, southern California (Fig. 1). It is a highly urbanized watershed with an area of about 400 km^2. As of 2001, around 70% of NBW was residential, commercial and industrial areas, and agricultural and orchard areas accounted for no more than 8%. It has a typical Mediterranean climate featured by short, mild winters, and dry summers. The annual rainfall depth is about 330 mm, occurring mostly between November and April. About 95% of the freshwater flow volume into the upper

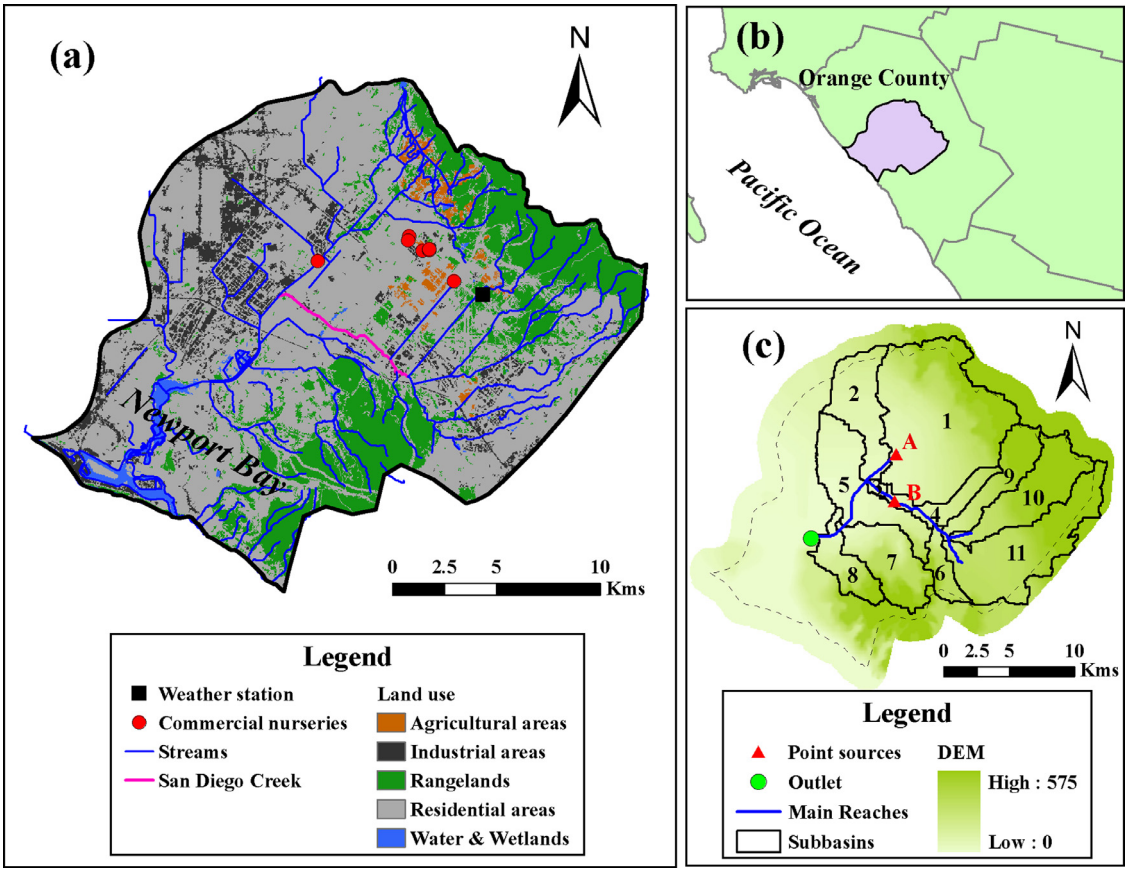

Fig. 1. The study area and modeling domain. (a) The Newport Bay Watershed (NBW); (b) location of NBW; and (c) watershed delineation in the SWAT model, where A and B denote two conceptualized point sources.

Newport Bay was delivered by San Diego Creek [5]. In 1990's, nutrients pollution was a key environmental issue in this area. The 1998 California 303(d) list of impaired waters for TMDL actions included San Diego Creek and Newport Bay, with nutrients as a major pollutant stressor. Commercial nurseries were the major point sources of nitrogen, and the effluent from these nurseries flows through channels (not illustrated) into the river network. Fertilizer application, urban stormwater and groundwater loading were the major non-point sources [39].

3.2. Streamflow and nitrate modeling using SWAT

SWAT is a widely used WWQ model developed by USDA-ARS and has been used in numerous applications [40–42]. It simulates hydrological processes, soil erosion, plant growth, and fate and transport of chemicals at a watershed scale, and considers many agriculture and water management practices. SWAT delineates a watershed into multiple sub-basins, each has a river reach to drain the runoff generated in it. A sub-basin can include multiple hydrologic response units (HRUs). HRU is the basic unit of model calculation, representing a specific combination of land use type, soil group and slope class. SWAT runs on a daily time-step. More details about the current version of SWAT can be found in Neitsch et al. [2].

To set up the SWAT model for the nitrate pollution at NBW, various types of data were collected. The 90 m Digital Elevation Model (DEM) was obtained from U.S. Geological Survey (USGS) National Elevation Dataset (NED). The 30 m land use raster data (year 2001) were collected from USGS National Land Cover Databases (NLCD). The soil data were achieved from U. S. State Soil Geographic (STATSGO) database. The meteorological observations (from 1997 to 2010) at Tustin-Irvine Ranch station (see Fig. 1a) were acquired from U. S. National Climatic Data Center (NCDC). NBW was delineated into 11 sub-basins (Fig. 1c) in the model. There are 58 HRUs in total, and 39 of them involve urban land uses (i.e., residential, commercial and industrial). In this study, streamflow and nitrate concentration at the outlet of the watershed (Fig. 1c) were the two model responses considered. Observations of daily flow and weekly nitrate concentration from July 2000 to June 2004 were collected from reports on the regional monitoring program established by the Santa Ana Regional Water Quality Control Board (SARWQCB).

The data on nitrate discharge from the major nurseries were collected from SARWQCB. In the SWAT model, the nurseries were conceptualized as two point sources (Fig. 1c), based on where the effluent entered the river. The nitrate loads at the sources A and B were estimated to be 61 kg/day and 31 kg/day, respectively, and assumed to be time-invariant because no temporal information was available. The estimated daily loads are highly uncertain in both their magnitude and variability. The model also simulates a common type of lawn grass, Bermuda grass (BERM in SWAT), in all the 39 urban HRUs. Fertilizer application in NBW was mainly for urban lawns. No detailed land application data were available for this area. Based on the information from University of California Cooperative Extension (www.ipm.ucdavis.edu/TOOLS/TURF/MAINTAIN/fertrate.html), the fertilizer usage in NBW was assumed to be 240 kg N/ha/year, and scheduled in six operations (40 kg N/ha each) of land application. In the SWAT model, the timing of the operations was determined by heat unit (HU) scheduling [2]. In NBW, urban stormwater runoff is another critical non-point source of nitrate. SWAT offers two options to simulate urban stormwater runoff pollution: one is the embedded USGS equations and the other is the buildup-washoff

Table 1
SWAT parameters investigated by the sensitivity analysis.

Category	Parameter	Definition	Range	Type
Hydrology parameters	ESCO	Soil evaporation compensation factor	[0.01, 1]	I
	EPCO	Plant evaporation compensation factor	[0.01, 1]	I
	CANMX	Maximum canopy index (mm H_2O)	[0.1, 10]	I
	SLSUBBSN	Average slope length (m)	[0.8, 1.2]	III
	HRU_SLP	Average slope steepness (m/m)	[0.8, 1.2]	III
	CN2	SCS runoff curve number for moisture condition II	[-20, 15]	II
	SURLAG	Surface runoff lag coefficient	[0.1, 12]	I
	OV_N	Manning's "n" value for overland flow	[-0.05, 0.1]	II
	SOL_AWC	Available water capacity of the soil layer (mm H_2O/mm soil)	[0.8, 1.2]	III
	SOL_K	Soil conductivity (mm/h)	[0.5, 1.5]	III
	CH_K1	Effective hydraulic conductivity in tributary channel alluvium (mm/h)	[1, 25]	I
	CH_K2	Effective hydraulic conductivity in main channel alluvium (mm/h)	[1, 25]	I
	CH_N1	Manning's "n" value for the tributary channels	[0.01, 0.2]	I
	CH_N2	Manning's "n" value for the main channel	[0.01, 0.2]	I
	TRNSRCH	Fraction of transmission loss from main channel to deep aquifer	[0, 0.1]	I
	ALPHA_BNK	Baseflow alpha factor for bank storage (days)	[0.001, 1]	I
	SHALLST	Initial depth of water in the shallow aquifer (mm H_2O)	[500, 2000]	I
	GW_REVAP	Groundwater 'revap' coefficient	[0.02, 0.2]	I
	REVAPMN	Threshold depth of water in the shallow aquifer for 'revap' (mm H_2O)	[0, 1000]	I
	GW_DELAY	Groundwater delay time (days)	[10, 800]	I
	ALPHA_BF	Baseflow recession factor (days)	[0.001, 1]	I
	RCHRG_DP	Groundwater recharge to deep aquifer	[0, 0.3]	I
	GWQMN	Threshold water depth in shallow aquifer for return flow to occur (mm H_2O)	[0, 1000]	I
Water quality parameters	NPERCO	Nitrogen percolation coefficient	[0.01, 1]	I
	SOL_NO_3	Initial NO_3 concentration in the soil layer (mg N/kg soil)	[0.5, 3]	III
	GW_NO_3	Nitrate concentration in the groundwater (mg N/L)	[0, 10]	I
	CDN	Denitrification exponential rate coefficient	[0.1, 3]	I
	SDNCO	Denitrification threshold water content	[0.2, 1.2]	I
	RSDCO	Residue decomposition coefficient	[0.01, 0.1]	I
	CMN	Rate factor for humus mineralization of active organic nutrients	[1e-4, 5e-4]	I
	N_UPDIS	Nitrogen uptake distribution parameter	[1, 30]	I
	BIOMIX	Biological mixing efficiency	[0, 1]	I

method [2]. Although the buildup-washoff method is conceptually appealing, we do not have local data to parameterize the equations. In the numerical experiments, the USGS option was chosen in the real-situation modeling scenario, while both options were considered in a synthetic modeling scenario, as will be explained in Section 3.5. Overall, the nitrate modeling has profound input and model structure uncertainty.

3.3. Sensitivity analysis

WWQ models like SWAT usually have 10s or even 100s of parameters. A sensitivity analysis (SA) is desired to choose a subset of critical parameters for calibration and uncertainty assessment. In this study, 32 SWAT parameters (Table 1), out of more than a hundred, were initially targeted. They are mathematically relevant for hydrology and nitrate simulations, and many of them were identified as sensitive ones in previous studies [40–44]. As most of the parameters have spatially distributed values, three strategies were considered in varying them, following Yang et al. [26]: varying the parameter value directly (Type I in Table 1); adding a deviation variable to the prior parameter value, and varying the deviation variable (Type II in Table 1); and applying a multiplier to the prior parameter value, and varying the multiplier (Type III in Table 1). All these random variables, either the original model parameters or the deviation and multiplier variables, were assumed to follow uniform distributions *a priori*, whose ranges are provided in Table 1. The proposed ranges were based on the SWAT's manual [45], as well as on references [40–44].

Many techniques are available for conducting a global SA, among which Morris screening [46,47] is an efficient and effective one, if sensitivity ranking, rather than quantification of variance contribution, is the aim. According to this method, given m uncertain parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_m)$, the parameter space can be discretized into a p-level grid Ω. At a given point in Ω, the elementary effect (EE) of the jth parameter is

$$EE_j = \frac{f(\theta_1, \ldots, \theta_j + \Delta, \ldots, \theta_m) - f(\theta_1, \ldots, \theta_j, \ldots, \theta_m)}{\Delta} \quad (6)$$

where $f(\bullet)$ represents concerned model outputs or certain goodness-of-fit measures, and Δ is a value in $\{1/(p-1), \ldots, 1\text{-}1/(p-1)\}$, such that the transformed point $(\theta_1, \ldots, \theta_j + \Delta, \ldots, \theta_m)$ is still in Ω. Morris screening produces multiple trajectories in Ω, each contains m+1 points (i.e., m+1 parameter realizations). Two consecutive points in a trajectory differ in only one component θ_q, and the difference is either Δ or $-\Delta$. In a trajectory, q is randomly picked from {1,2, …,m} in each move, with no duplication. If the number of trajectories is r, r EEs can be computed for each parameter with $r(m+1)$ model evaluations. The mean of $|EE|$ is usually a preferred sensitivity index (SI), for it avoids the problem that EEs with opposite signs may cancel out when the model is non-monotonic [46].

This study adopted the Morris screening method to select parameters for the Bayesian inference. The level p was set to 6 and Δ was fixed at $p/(2(p-1)) = 3/5$ to guarantee that each level has equal probability of being selected. The Nash–Sutcliffe efficiency (NSE) coefficient [48] was a goodness-of-fit measure, and considered as the $f(\bullet)$ in Eq. 6. The number of trajectories (i.e., r) was 200, and therefore 6,600 (= 200*(32+1)) evaluations of the SWAT model were carried out to complete the SA. The mean of $|EE|$ was used as the sensitivity index (SI) to rank the parameters. As mentioned before, streamflow and nitrate concentration at the watershed outlet were the two model responses considered in this study. Thus, the SI was evaluated with regard to both responses. Naturally, the SI ranking varies for the two responses. Table 2 lists the 12 most important parameters identified by the SA, each of which ranks top 8 (i.e., top 25%) among the 32 parameters at least for one of the two responses. These 12 parameters were considered as random variables in the Bayesian inference.

Table 2
Uncertain parameters inferred by the Bayesian calibration.

Parameter	SI rank for flow	SI rank for NO_3	Synthetic *true* value	Unit
SURLAG	1	3	2.235	/
CH_K1	2	6	1.458	mm/h
GW_DELAY	3	8	723.69	day
CN2	4	22	–9.770	/
CH_K2	5	23	7.094	mm/h
ALPHA_BNK	6	24	0.0047	day
CH_N1	7	16	0.0511	/
ESCO	8	2	0.0232	/
NPERCO	28	1	0.0843	/
SDNCO	24	4	0.2556	/
SHALLST	12	5	644.60	mm H_2O
SOL_AWC	9	7	0.1103	mm H_2O/mm soil

3.4. Bayesian inference using DREAM$_{(ZS)}$

In the field of hydrologic modeling, Markov chain Monte Carlo (MCMC) is now the most popular sampling scheme for Bayesian inference. In this study, DREAM$_{(ZS)}$, a state-of-the-art MCMC algorithm, was employed to perform the multiple-response Bayesian calibration. Details about this algorithm can be found elsewhere [33,34,49]. In brief, DREAM$_{(ZS)}$ is a variant of DREAM. Like the original DREAM algorithm, DREAM$_{(ZS)}$ runs multiple Markov chains simultaneously to facilitate the global exploration of parameter space, and automatically tunes the scale and orientation of the proposal distribution by using randomized subspace sampling. DREAM$_{(ZS)}$ further reduces the required number of parallel chains by generating trial moves based on an archive of past states. It also includes a snooker updater [49] to increase the diversity of candidate points.

In this study, the hyper-parameters (i.e., $\boldsymbol{\varphi}$) in the combined likelihood function were not jointly inferred with the uncertain SWAT parameters (i.e., $\boldsymbol{\theta}$). In one MCMC sampling step, a realization of the SWAT parameters was achieved first; and given this realization, the corresponding values of the hyper-parameters were then determined by maximum likelihood estimation (MLE), as suggested by Hantush and Chaudhary [50]. Here we use $\hat{\boldsymbol{\varphi}}$ to denote the corresponding MLE result of $\boldsymbol{\varphi}$. The classic SCE-UA algorithm [51] was used to implement the MLE. The MLE was performed for different realizations of $\boldsymbol{\theta}$ to achieve the corresponding $\hat{\boldsymbol{\varphi}}$ vectors. In this study, these $\hat{\boldsymbol{\varphi}}$ vectors were used to quantify the predictive uncertainty, although they should not be considered as an approximation of the posterior distribution of $\boldsymbol{\varphi}$. As explained in Appendix A, the mathematical nature of this MLE-based inference is to sample realizations of $\boldsymbol{\theta}$ from the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{Z})$, which is approximated by $p(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}|\mathbf{Z})$, instead of the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z})$. Substituting $p(\boldsymbol{\theta}|\mathbf{Z})$ for $p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z})$ in the sampling would not alter the posterior distribution of $\boldsymbol{\theta}$, but could significantly improve the MCMC sampling efficiency, because the dimension of sampling space is reduced. In fact, preliminary tests showed that this MLE-based inference enhanced the acceptance rate of Markov chain several folds, compared to the regular joint inference. On the other hand, as $\hat{\boldsymbol{\varphi}}$ is at the center of the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})$ (refer to the Appendix), the distribution of $\hat{\boldsymbol{\varphi}}$ should have the same central tendency as, but narrower than, the posterior distribution of $\boldsymbol{\varphi}$. As a quality control measure, the MLE-based approach was compared against the regular approach for our modeling case. Figs. S6 to S9 in the Supplementary Materials show an example of the comparison results. The figures clearly show that the two approaches led to almost identical posterior distributions of the SWAT parameters (Fig. S6) and uncertainty results (Figs. S8 and S9), and the difference between the distribution of $\hat{\boldsymbol{\varphi}}$ and the posterior distribution of $\boldsymbol{\varphi}$ is consistent with the theory (Fig. S7).

The *R* metric suggested by Gelman and Rubin [52] was used to evaluate the convergence of Markov chains. Parallel Markov chains were deemed to achieve convergence when the *R* values for all random parameters were constantly below the threshold of 1.2 in last 20,000 points. The last 50% points of the Markov chains were then used for subsequent uncertainty analysis. In this study, 12 Markov chains were adopted in DREAM$_{(ZS)}$, and 12 CPUs (2.67 GHz) were used in parallel to reduce the computing time. It is worth pointing out that, for our specific modeling case, there could be more efficient sampling algorithms than DREAM$_{(ZS)}$. We used the original DREAM$_{(ZS)}$ algorithm (except that the MLE inference was embedded) nevertheless, because the computational cost of implementing DREAM$_{(ZS)}$ was affordable, and our study was not aimed to optimize the MCMC sampling.

3.5. Numerical experiments

The numerical experiments were based on both the real-situation modeling scenario (see Section 3.2) and a synthetic modeling scenario (Table 3) introduced below. In this synthetic scenario, *true* values of the top 12 sensitive parameters were first hypothesized, as listed in the fourth column of Table 2. This synthetic parameter set was picked from a number of randomly generated realizations, and ensures that the corresponding flow and nitrate simulations are reasonable and representative. The remaining SWAT parameters took their model default values. *True* nitrate source loadings for calibration and validation periods were assumed as well (the second column in Table 3). With the hypothetic *true* parameter values and loading inputs, the *true* model responses of flow and nitrate concentration at the watershed outlet were achieved. The synthetic scenario incorporates three modeling cases (Table 3), a *true* (or perfect) model with known *true* parameter values, and two imperfect models whose uncertain parameters required to be inferred. Hereafter, the *true* and imperfect models were abbreviated as TM, IM1 and IM2, respectively. IM1 was assumed to have no input and structure errors, and identical

Table 3
Three hypothetic SWAT models in the synthetic modeling scenario.

Setting	*True* (or perfect) model (TM)	Imperfect model w/o input & structure errors (IM1)	Imperfect model with input & structure errors (IM2)
True parameter values	Known	Unknown; to be inferred	Unknown; to be inferred
Calibration period (CP)	Jul. 2000 – Jun. 2004	Jul. 2000 – Jun. 2004	Jul. 2000 – Jun. 2004
Point source A in CP[a]	132 kg N/day (dynamic)	132 kg N/day (dynamic)	100 kg N/day (constant)
Point source B in CP[a]	68 kg N/day (dynamic)	68 kg N/day (dynamic)	50 kg N/day (constant)
Fertilizer use in CP[b]	240 kg N/ha/year (Dates)	240 kg N/ha/year (Dates)	165 kg N/ha/year (HU)
Stormwater Model in CP	Buildup-washoff	Buildup-washoff	USGS equations
Validation period (VP)	Jul. 2004 – Jun. 2008	Jul. 2004 – Jun. 2008	/
Point source A in VP[a]	140 kg N/day (dynamic)	140 kg N/day (dynamic)	/
Point source B in VP[a]	74 kg N/day (dynamic)	74 kg N/day (dynamic)	/
Fertilizer use in VP[b]	240 kg N/ha/year (Dates)	240 kg N/ha/year (Dates)	/

[a] "Dynamic" means that the daily loading input into the model is time-variant, while "constant" indicates constant daily loading.
[b] "Dates" means that the land application operations were scheduled on specific dates, while "HU" indicates the heat unit scheduling approach.

Table 4
Future loading plans in synthetic and real-situation modeling scenarios.

Future loading	Synthetic modeling	Real-situation modeling
Plan 1 (increased loading)		
Point source A[a]	140 kg N/day	90 kg N/day
Point source B[a]	70 kg N/day	45 kg N/day
Fertilizer use[b]	248 kg N/ha/year	356 kg N/ha/year
Plan 2 (steady loading)		
Point source A[a]	100 kg N/day	60 kg N/day
Point source B[a]	50 kg N/day	30 kg N/day
Fertilizer use[b]	165 kg N/ha/year	238 kg N/ha/year
Plan 3 (reduced loading)		
Point source A[a]	60 kg N/day	30 kg N/day
Point source B[a]	30 kg N/day	15 kg N/day
Fertilizer use[b]	83 kg N/ha/year	119 kg N/ha/year

[a] Constant daily loading.
[b] Land application operations were scheduled on specific dates.

Table 5
Convergence speeds of Bayesian calibration in different experiments.

	Model	One-response calibration	Two-response calibration
Number of model runs (CPU time)	IM1	54,720 (3.8 days)	115,200 (8.0 days)
	IM2	65,760 (4.6 days)	150,000 (10.4 days)
	RM	49,920 (3.5 days)	117,120 (8.1 days)
Acceptance rate[a]	IM1	10.09%	4.20%
	IM2	12.94%	4.59%
	RM	13.26%	3.74%

[a] Acceptance rate is the probability of candidate points that are accepted by the Markov chains in stationarity.

to TM except that its parameter values are unknown; while IM2 was assumed to have both model input errors (i.e., incorrect point and non-point sources loadings) and model structure errors (i.e., using the USGS option instead of the buildup-washoff option). IM1 represents a situation in which only parametric uncertainty presents, while IM2 reflects a common situation that input and model structure errors are significant, in addition to the parametric uncertainty. Fig. S1 in the Supplementary Materials illustrates the synthetic *true* responses for the calibration period.

To perform the Bayesian calibration for IM1 and IM2, daily observations of flow and weekly observations of nitrate concentration were also synthesized by corrupting the *true* model responses with hypothetic observational errors. The hypothesized errors were generated from non-autocorrelated, heteroscedastic, and non-Gaussian error models with $\sigma_0^H = 0.2$, $\sigma_1^H = 0.05$, $\beta^H = 0.25$, $\xi^H = 3$, $\phi^H = 0$, $\sigma_0^N = 0.05$, $\sigma_1^N = 0.2$, $\beta^N = 0.5$, $\xi^N = 0.9$ and $\phi^N = 0$, where the superscripts H and N represent flow and nitrate responses, respectively. Fig. S1 in the Supplementary Materials illustrates the synthetic observations for the calibration period.

Both single-response (i.e., flow) and two-response (i.e., flow and nitrate concentration) Bayesian calibrations were performed for the two imperfect models, IM1 (with parametric uncertainty only) and IM2, as well as the real-situation model (denoted as RM) introduced in Section 3.2, for the period between July 2000 and June 2004. Note that, in each SWAT run, a two-year (07/01/1998–06/31/2000) warm-up period was added prior to the calibration period. Thus, the total simulation length in the calibration stage was six years. Water quality monitoring data are often sparse in time. This study dealt with weekly nitrate observations. For IM1, the same type of error models for synthesizing the observations was used. For IM2 and RM, we gradually increased the complexity of error models, from simple Gaussian, homoscedastic and non-autocorrelated type to the most complicated type, until posterior checks confirmed that the residual errors are consistent with the error model assumptions. Posterior justification of these error assumptions are illustrated by Figs. S2–S5 in the Supplementary Materials. Eventually, for IM2, we adopted a non-Gaussian, heteroscedastic and autocorrelated error model for the flow response and a non-Gaussian, heteroscedastic and non-autocorrelated error model for the nitrate response. The same types of error models were considered for RM, except that homoscedasticity is adopted for the nitrate response. Therefore, there are nine hyper-parameters in total in the combined likelihood function for IM2, and eight for RM. The prior ranges of the hyper-parameters were set to be $\sigma_0^H \in [0, \sigma_H]$, $\sigma_0^N \in [0, \sigma_N]$, $\sigma_1^H \in [0, 1]$, β^H, $\beta^N \in [-1, 1]$, ξ^H, $\xi^N \in [0.1, 10]$, $\phi^H \in [0, 1]$ for both IM2 and RM, and $\sigma_1^N \in [0, 1]$ for IM2. Here σ_H and σ_N are the standard deviations of the observed flow and nitrate concentration, respectively.

To further compare the one-response Bayesian calibration and the two-response Bayesian calibration, hereafter respectively denoted as 1R-BC and 2R-BC, a four-year validation period (VP) between July 2004 and June 2008 was considered. *True* source loadings were assumed for this period (Table 3) to derive the *true* model responses by TM (see Fig. S10 in the Supplementary Materials), as well as the stochastic simulations by IM1. The artificial observations for this period (see Fig. S10 in the Supplementary Materials) were generated from the synthetic *true* responses in the same way as for the calibration period. To investigate how the different calibration strategies would impact management decisions, three future loading plans were proposed (the second column in Table 4), and then evaluated

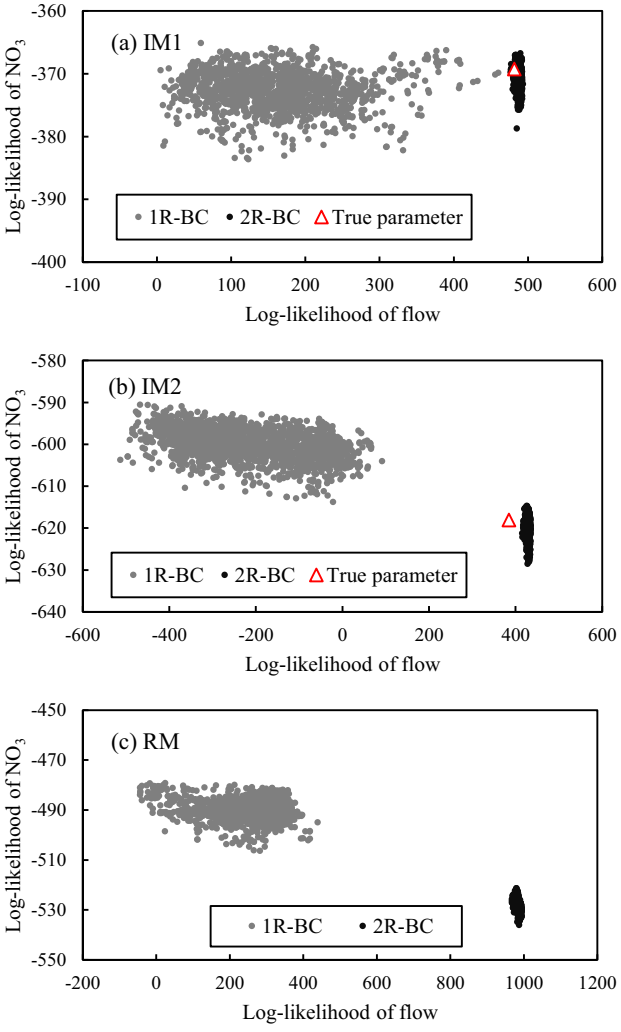

Fig. 2. Log-likelihood values of flow and nitrate concentration derived in both one-response calibration (1R-BC) and two-response calibration (2R-BC). (a) the synthetic imperfect model without input and structure errors; (b) the synthetic imperfect model with input and structure errors; and (c) the real-situation model.

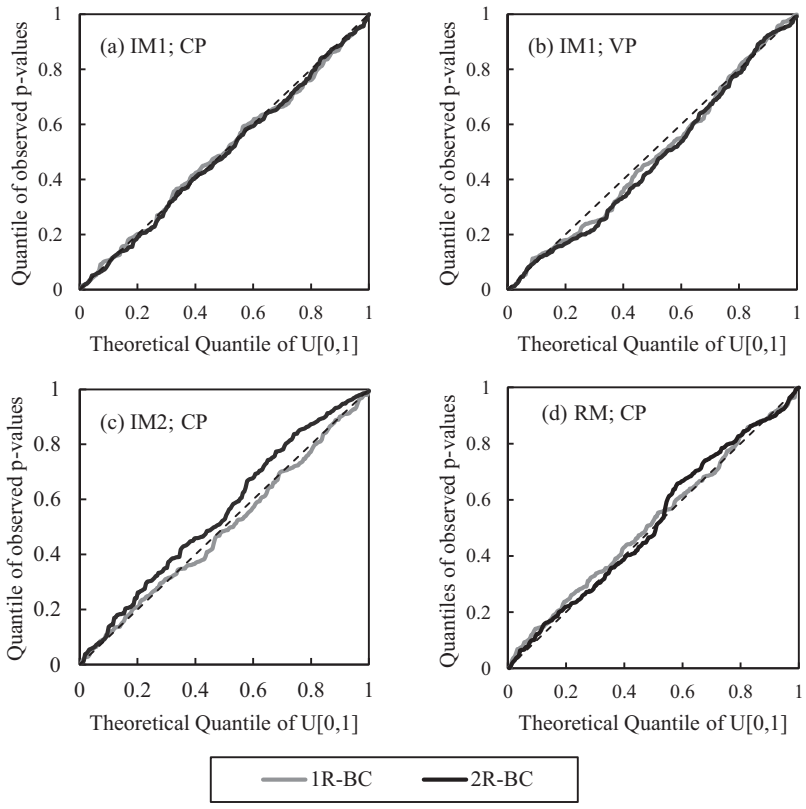

Fig 3. QQ plots of nitrate concentration derived from both one-response calibration (1R-BC) and two-response calibration (2R-BC). (a) IM1 in the calibration period (CP); (b) IM1 in the validation period (VP); (c) IM2 in CP; and (d) RM in CP.

deterministically by TM and stochastically by IM2. Compared to the loading in the calibration period (the fourth column in Table 3), they represent increased, steady and decreased loadings, respectively. Similarly, three future loading plans were proposed in the real-situation modeling case as well (the third column in Table 4).

In the Bayesian inference for IM1 and IM2, the *true* values of those non-random parameters were set to be their default values. In reality, the *true* values would be unknown, and where to fix them inevitably involves subjectiveness. In the Bayesian inference for RM, a pre-calibration was performed for the 20 parameters (including the 12 parameters in Table 2) whose sensitivity ranks top 16 with regard to the flow and/or nitrate concentration. The pre-calibration was accomplished using SCE-UA (10,000 SWAT evaluations in maximum). The log-likelihood functions, L^N and $L_{multiple}$, were used as the objective functions for 1R-BC and 2R-BC, respectively. In either case, the 8 non-random ones in the 20 parameters were then fixed at their pre-calibrated values in the Bayesian inference, as suggested by our previous study [53]. All other parameters (with no significant impact on the model simulations) were fixed at their default values. All the experiments and data analyses were programmed using MATLAB.

4. Results and discussions

4.1. Convergence speed and effectiveness of calibration

Computational cost is a great concern for MCMC-based Bayesian calibration. Although the SWAT models in this study only cost 5 to 6 s to complete a six-years simulation on a good-performance CPU core (2.67 GHz), 10,000 model evaluations still require about a half day to finish. The convergence speeds of the DREAM$_{(ZS)}$ algorithm in different numerical experiments were compared in Table 5. Although costly, the calibration is still feasible with 12 CPU cores used in parallel. For the most expensive experiment (i.e., 2R-BC for RM), the actual computing time was about one day. It has also been noticed that adding the flow response in the calibration reduced the acceptance rate and slowed down the convergence of DREAM$_{(ZS)}$. This is because the posterior parameter distributions would be more complex when more responses are considered in the inference.

Fig. 2 plots the log-likelihood values of the last 50% of the MCMC samples in each calibration experiment, as well as the values of the hypothesized *true* parameter set. Note that in one-response Bayesian calibration (1R-BC), the log-likelihood of flow was not evaluated during the MCMC sampling, but determined after all the samples were achieved. Some important observations were obtained, as discussed below.

First of all, for IM1 with no input and model structure errors (Fig. 2a), adding the flow response significantly enhanced the log-likelihood of flow (*x*-axis), which is self-evident, while slightly improved the log-likelihood of nitrate concentration (*y*-axis). The log-likelihood values of the *true* parameter set were captured by those derived by two-response Bayesian calibration (2R-BC), indicating the effectiveness of the Bayesian calibration using DREAM$_{(ZS)}$.

Second, the *true* parameter set has different log-likelihood values in the modeling cases IM1 and IM2. With the introduced input and structure errors in IM2, the log-likelihood value for flow was mildly reduced from 481.20 (Fig. 2a) to 384.13 (Fig. 2b), but the log-likelihood value for nitrate concentration was abruptly reduced from −369.35 (Fig. 2a) to −625.40 (Fig. 2b). It appears that the introduced errors have caused much more significant uncertainty for the water quality simulation than for the flow simulation, which represents a very common situation in reality.

Third, for IM2, introducing the flow response enhanced the log-likelihood of flow but substantially degraded that of nitrate

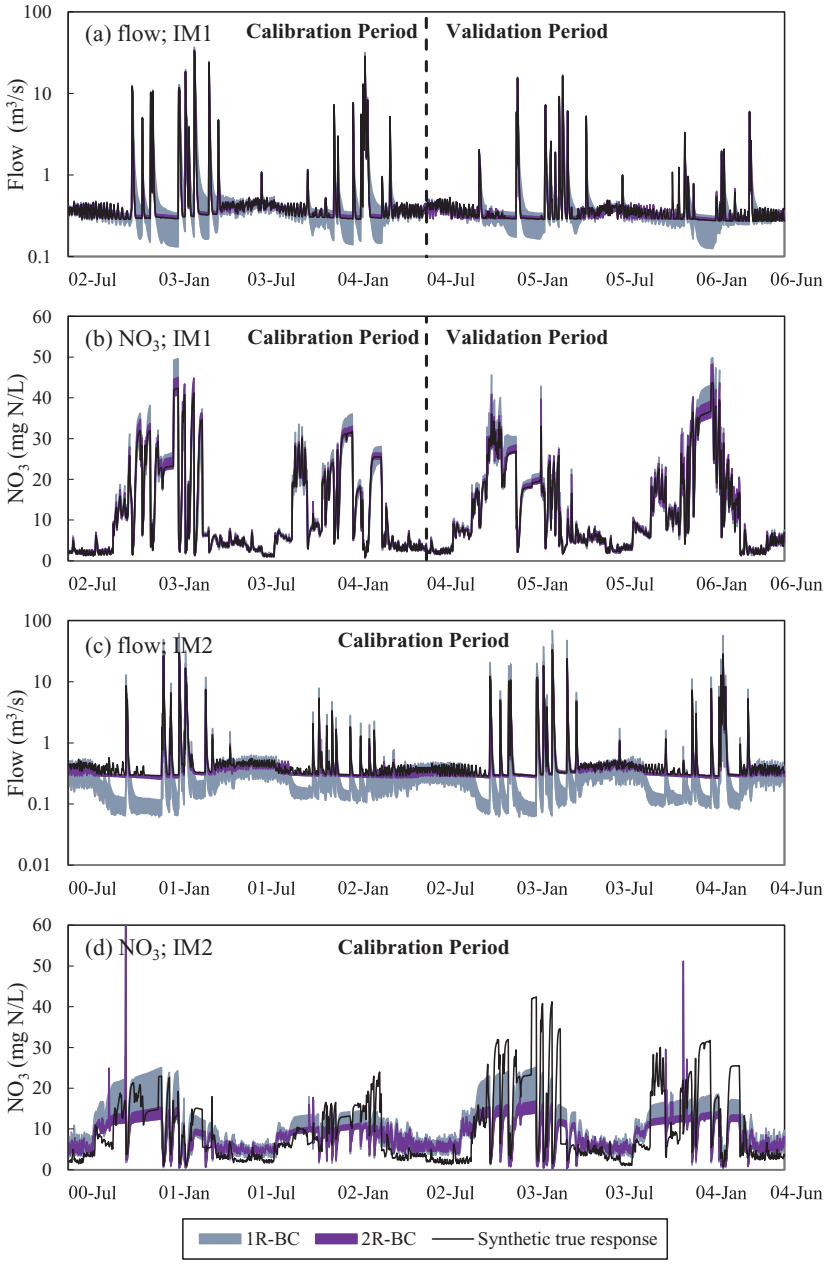

Fig. 4. Parametric uncertainty bands of flow and nitrate concentration. (a) flow simulated by IM1; (b) nitrate concentration simulated by IM1; (c) flow simulated by IM2; and (d) nitrate concentration simulated by IM2.

concentration at the meantime (Fig. 2b). This notable tradeoff can be explained as follows. As the uncertainty in flow modeling due to the input and structure errors is relatively small, the 2R-BC with flow observations incorporated tends to adequately capture the *true* values of those flow-related parameters. As the input and structure errors result in larger uncertainty in nitrate modeling, better capture of the *true* flow-related parameter values would reduce the extent to which parametric uncertainty may compensate the input and model structure error with regard to nitrate modeling. Nevertheless, it is worth pointing out that the reduced compensation due to adding flow observations is not necessarily a negative effect, which will be further discussed in Sections 4.3 and 4.4.

In addition, the likelihood values in the real-situation model (RM) case (Fig. 2c) exhibit the same pattern as IM2 (Fig. 2b). More discussion on RM will be provided in Section 4.5.

4.2. Uncertainty results

Fig. 3 shows the predictive quantile-quantile (QQ) plots of nitrate concentration. The QQ plots were derived based on observed *p*-values, which are the cumulated probabilities of observations in the predictive distribution [54]. A QQ plot can provide intuitive information on consistency of predictive distribution with observations [54]. In a QQ plot, if all the points (i.e., the curve in Fig. 3) are on the 45° line, the predictive distribution agrees perfectly with the observations. As Fig. 3(a and b) indicate, for IM1 (i.e., the synthetic imperfect model without input and structure errors), the predictive distributions of nitrate concentration derived by 1R-BC and 2R-BC can well represent the observed variability in both calibration and validation periods. For IM2 (i.e., the imperfect model with input and structure errors), the predictive distribution by 1R-BC stays close to the 45° line,

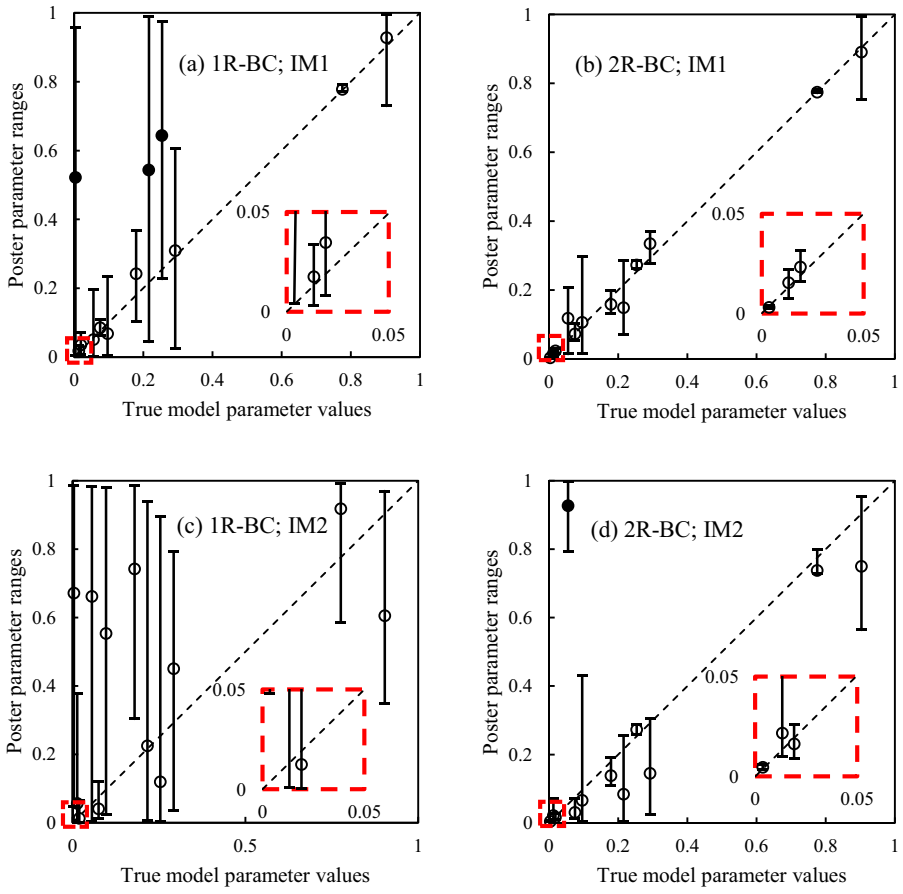

Fig. 5. Posterior distributions of the random parameters. The circles are the medians, the upper bars are the 97.5 percentiles, and the lower bars are the 2.5 percentiles. The dashed square in the lower-right corner of each subplot presents a zoom-in view of the near-origin zone. (a) 1R-BC for IM1; (b) 2R-BC for IM1; (c) 1R-BC for IM2; and (d) 2R-BC for IM2.

but the distribution by 2R-BC has a notable deviation (Fig. 3c). It is consistent with Fig. 2b, and again suggests that the input and model structure errors have been largely uncompensated by the parametric uncertainty with regard to nitrate modeling, due to the incorporation of flow response in the calibration.

Fig. 4 illustrates the parametric uncertainty in different modeling cases. The bands represent the 95% uncertainty intervals based on the last 50% of the MCMC samples. In general, 2R-BC led to much narrower uncertainty bands than 1R-BC, reflecting the additional constraining effect of the flow observations. For IM1 (Fig. 4a and b), the uncertainty bands of both flow and nitrate concentration produced by 2R-BC precisely and tightly embrace the respective *true* responses. Although wider, the bands of 1R-BC still well capture the respective *true* responses in this ideal no-error case. An important implication drawn by comparing Fig. 4b with Fig. 3a and b is that, although 1R-BC and 2R-BC do not differ much in the consistency of predictive distribution with observations, they could have very different parametric uncertainty results.

For IM2, the uncertainty band of flow by 1R-BC is seriously biased (Fig. 4c). The bias is towards the low flows, because the flow observations were not assimilated and the *true* pollutant loadings were underestimated. On the other hand, the bands of nitrate concentration by 1R-BC and 2R-BC both deviate from the *true* response (Fig. 4d). The bias is towards the high concentrations in certain periods, while towards the low concentrations in others. It is mainly because the point source loadings fed to IM2 were time-invariant, while the *true* ones have temporal variability. The band by 2R-BC is even worse than the one by 1R-BC, as the former largely misses those concentration peaks. It actually indicates the reduced compensation effect of parametric uncertainty due to the additional flow observations. As mentioned before, Fig. 4d does not mean that 2R-BC is worse than 1R-BC, as will be demonstrated in Sections 4.3 and 4.4.

4.3. Posterior parameter distributions

Fig. 5 further compares the posterior parameter distributions derived by one-response Bayesian calibration (1R-BC) and two-response Bayesian calibration (2R-BC). All the parameter values were rescaled to [0, 1] by a linear min–max normalization. The circles represent medians, the upper bars are 97.5 percentiles, and the lower bars are 2.5 percentiles. A circle on the diagonal line suggests that the respective posterior distribution is centered at the *true* value (i.e., with no bias), and a narrower bar interval indicates a better identified posterior distribution. In the case of 1R-BC for IM1 (Fig. 5a), three parameters (with solid circles) have biased posterior distributions and they are CH_N1 (Manning's *n* for the tributary channels), CH_K2 (effective hydraulic conductivity in main channel alluvium) and ALPHA_BNK (baseflow alpha factor for bank storage). Their posterior ranges are very large as well, indicating that they have been poorly identified. It is not surprising because they are all hydrology parameters to which the nitrate response is not sensitive (see Table 2). On the contrary, Fig. 5b shows that, with the flow response considered, all the parameters have been adequately identified.

For IM2 with both input and model structure errors, almost all of the parameters were poorly identified by 1R-BC (Fig. 5c), because the significant input and structure errors led to large variance of residual errors and high parametric uncertainty bands (see Fig. 4d). The bias is also notable for most of the parameters, because the parametric uncertainty was compensating the input and structure errors with regard to nitrate modeling during the calibration. On the other

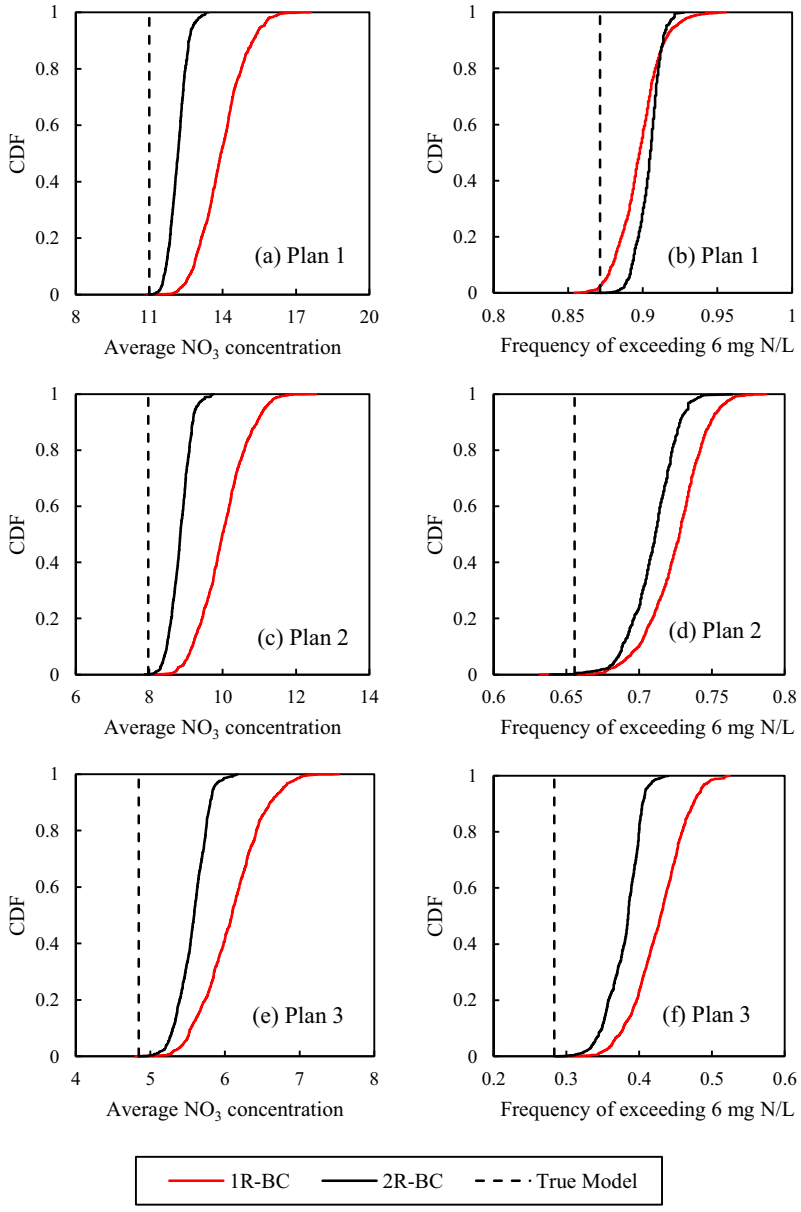

Fig. 6. Simulated management-relevant variables for three future loading plans using the *true* model and the imperfect model IM2. (a), (c) and (e) are for the average nitrate concentration; and (b), (d) and (f) are for the frequency of exceeding 6 mg N/L.

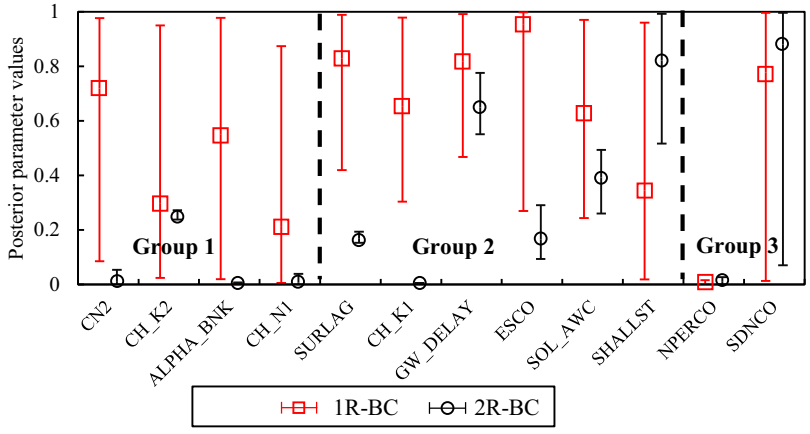

Fig. 7. Posterior distributions of random parameters derived by 1R-BC and 2R-BC, considering the real-situation model. The squares and circles are medians, the upper bars are 97.5 percentiles, and the lower bars are 2.5 percentiles.

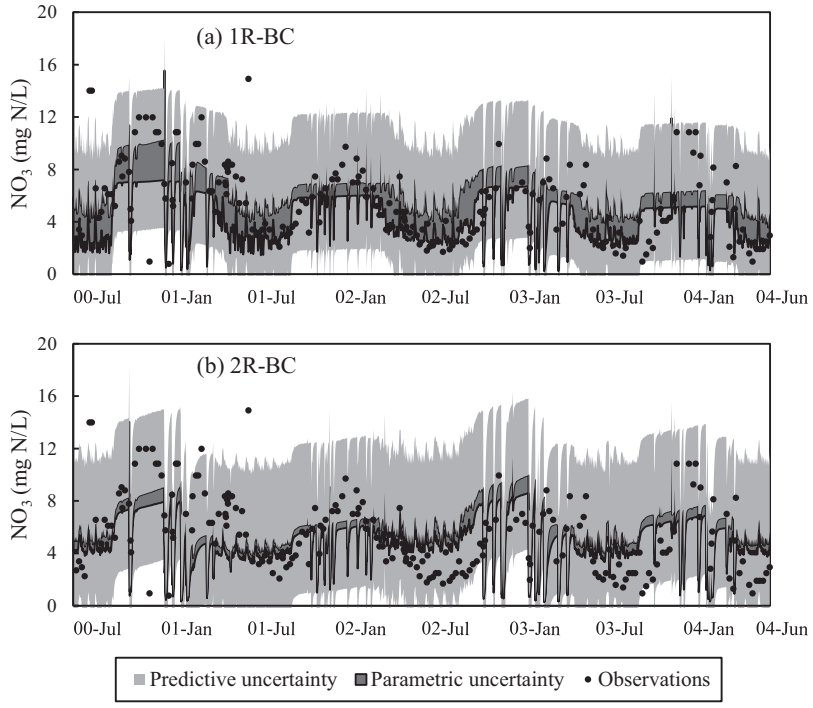

Fig. 8. Predictive and parametric uncertainty bands (95% confidence intervals) of nitrate concentration modeled by the real-situation model (RM) for the calibration period. (a) one-response Bayesian calibration; and (b) two-response Bayesian calibration.

hand, the posterior distributions derived in 2R-BC are much more reasonable (Fig. 5d), because the flow observations being assimilated helped drive the posterior distributions towards the respective *true* values. In Fig. 5d, the parameter with a solid circle is SDNCO, whose posterior distribution is dramatically off the *true* value. It is because the flow response is very insensitive to SDNCO (see Table 2), and the flow observations would not help much in identifying SDNCO's posterior distribution. Apparently, including the flow response that is less susceptible to the input and model structure errors actually improved the calibration results, although the predictive uncertainty (Fig. 3c) and parametric uncertainty (Fig. 4d) results were somewhat degraded.

In addition, the distributions of the hyper-parameters' MLE values (i.e., $\hat{\boldsymbol{\varphi}}$) for IM1 and IM2 are illustrated by Figs. S11 and S12 in the Supplementary Materials. In case of IM1, the distributions well embrace the respective synthetic *true* values.

4.4. Impact on management decisions

It has been shown so far that, if water quality simulation is susceptible to input and model structure errors, Bayesian calibration considering water quality response *only* would lead to unreliable calibration results, as the Bayesian inference tends to compensate input and structure errors with parametric uncertainty. By incorporating flow response which is usually much less impacted by the input and structure errors, the calibration results would be greatly improved. To further investigate the impact of input and model structure errors in a water quality management context, the three future loading plans in the synthetic modeling scenario (the second column in Table 4) were evaluated using the *true* model (TM), as well as using the last 50% of the MCMC realizations of IM2 (with both input and model structure errors) obtained in both 1R-BC and 2R-BC. As examples, two management-relevant variables, average nitrate concentration and frequency of exceeding the water quality target (i.e., 6mg N/L), were evaluated based on the simulations by TM and IM2.

The Cumulative Distribution Functions (CDFs) in Fig. 6 show that, for both of the management variables, the synthetic input and model structure errors caused an over-prediction under all the future loading conditions. However, in most of the cases, including the flow response alleviated the over-prediction. It confirms that 2R-BC is more appropriate than 1R-BC in the management context. Significant input and structure errors would inevitably bias modeling-based water quality management decisions. There are two potential solutions to this: one is to address the input and structure errors directly in the Bayesian framework, which is mathematically challenging and deserve further studies, because explicitly modeling these errors are often hard to be validated; and the other is to include additional responses that are less vulnerable to input and structure errors into a multiple-response calibration. Prior judgment on whether a model response is vulnerable to input and model structure errors would be then critical. The instream flow investigated in this study would be a general choice.

4.5. The real-situation model

The Bayesian calibration was also performed for the real-situation model (RM) introduced in Section 3.2. It is reasonable to assume that RM has significant input and model structure errors with regard to nitrate modeling, because the nitrate source loading data fed to the model are very rough estimates, and the urban runoff quality module highly simplifies real processes. The uncertainty in flow modeling due to input and structure errors is assumed to be much more limited, as NBW is a relatively small and flat watershed and the meteorological data we collected are of good quality. The assumptions are also supported by the similarity between the log-likelihood plot for IM2 (Fig. 2b) and that for RM (Fig. 2c).

Fig. 3d shows the QQ plots of nitrate concentration for RM, which suggest that the predictive distributions achieved by 1R-BC and 2R-BC are both acceptable. Fig. 7 illustrates the posterior parameter distributions derived by 1R-BC (in red) and 2R-BC (in black). The numbers of SWAT runs to reach the convergence in 1R-BC and

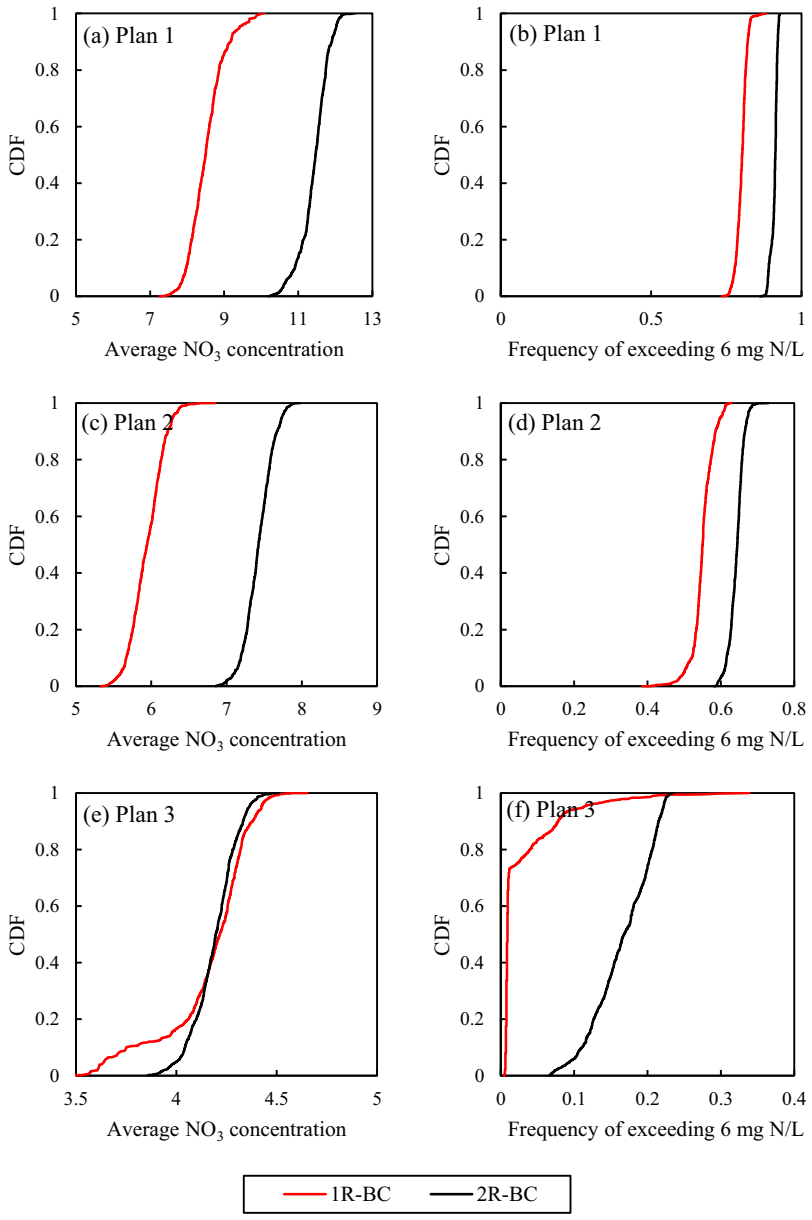

Fig. 9. Management-relevant variables evaluated by the real-situation model under the three future loading conditions. (a), (c) and (e) are for average nitrate concentration; and (b), (d) and (f) are for frequency of exceeding 6 mg N/L.

2R-BC were 49,920 and 117,120, respectively. Three distinctive groups can be identified: sensitive parameters for the flow simulation only (Group 1), sensitive parameters for both the flow and nitrate simulations (Group 2) and sensitive parameters for the nitrate simulation only (Group 3). As expected, by incorporating the flow response, the identifiability (reflected by the interval width) of Group 1 was significantly enhanced, while the identifiability of Group 3 was unchanged. In Group 2, the identifiability improvement increases with the parameter's flow sensitivity (see Table 2). In addition, for the two most flow-sensitive parameters, SURLAG and CH_K1, 1R-BC and 2R-BC ended up in very different regions of the parameter space. In this real-situation case, we are not able to directly prove that the posterior distributions of the random parameters were improved towards a right direction by 2R-BC, because we do not know their *true* values. Nevertheless, given what we observed in the synthetic modeling scenario (Fig. 5), it is reasonable to believe that the posterior distributions achieved by 2R-BC are probably less biased than those by 1R-BC.

Fig. 8 shows the stochastic simulation results of nitrate concentration by RM, using the last 50% of the MCMC samples obtained by both 1R-BC and 2R-BC. The discrepancy between the predictive uncertainty band and the respective parametric uncertainty band was characterized by the residual error model. The predictive uncertainty band was derived using the simulations by the last 50% of the MCMC samples (N_1). For each of the N_1 simulations, N_2 perturbed simulations were randomly generated based on the statistical error model. Then the 2.5 and 97.5 percentiles of the $N_1 \times N_2$ realizations were used to derive the predictive uncertainty limits. N_2 is set to be 1 by default [33,34]. In our study, we set N_2 to 100 to ensure robust results. The predictive uncertainty bands by 1R-BC and 2R-BC respectively embrace 97.6% and 96.7% of the observational data, which is a reasonable outcome. The wide bands imply the significant impact of the input and model structure errors. As for the parametric uncertainty, 2R-BC resulted in a narrower band than 1R-BC, reflecting the additional constraining effect of the flow observations. Both parametric uncertainty bands well reproduce the seasonal pattern of the

observations, that is, high nitrate concentrations in winter storm seasons and low concentrations in summer dry periods.

Stochastic simulations by RM were also performed for the three future loading plans (the third column in Table 4). The two management-relevant variables were evaluated based on the stochastic simulations, and their CDFs were plotted in Fig. 9. In general, 2R-BC resulted in higher average concentrations and exceedance frequencies, and the differences are notable. It implies that incorporating a model response less sensitive to input and model structure errors in the calibration stage can significantly improve the reliability of the model prediction and possibly lead to better management decisions. Interestingly, as demonstrated by Fig. 9, the underestimation by 1R-BC is more prominent for Plans 1 and 2 than for Plan 3 with regard to the average concentration; but more prominent for Plan 3 with regard to frequency of exceeding 6 mg N/L. On the other hand, the discrepancy between 1R-BC and 2R-BC varies notably across different confidence levels. It is most evident in Fig. 9f, in which the notable underestimation turns into a large overestimation at confidence levels higher than 0.99. Essentially, to what extent an additional response with less input and model structure errors would improve modeling-based decision-making actually depends on many factors including pollution severity, management objective and decision maker's risk tolerance.

5. Conclusions

This study implemented multiple-response Bayesian calibration (MRBC) to SWAT, a classic watershed water quality (WWQ) model, using the nitrate pollution in the Newport Bay Watershed as the case study. The impact of input and model structure errors on uncertainty quantification, parameter identification and management decision-making was systematically investigated. Major conclusions include the following. First, with an efficient MCMC algorithm, MRBC is applicable to WWQ modeling in characterizing its parametric and predictive uncertainties. The computational cost of MRBC is relatively high for WWQ modeling, but still affordable. Second, significant input and model structure errors would degrade the calibration of uncertain model parameters. But the degradation can be alleviated by including additional model responses that are less impacted by input and model structure errors in a MRBC framework. Particularly, instream flow observations of high frequency and good quality can greatly help the calibration. Third, significant input and model structure errors would inevitably mislead modeling-based management decisions. To what extent an additional response can improve the decision-making depends on pollution severity, management objective and decision maker's risk tolerance.

Overall, this study demonstrated the feasibility and benefits of performing MRBC for complex WWQ models within a management context, and demonstrated that appropriate interpretation of uncertainty in water quality modeling depends on management concerns. Future studies may address the following important issues. First of all, how to explicitly account for the input and model structure errors in the Bayesian calibration remains to be a major challenge. Second, evaluating a complex WWQ model for 100s of 1,000s of times is impractical in most real-world management applications, and low-cost surrogate models like Polynomial Chaos Expansion (PCE) [55] and Support Vector Machine (SVM) [56] could be adopted to replace the original model in the uncertainty analysis. Finally, software development efforts are needed to enhance the usability of MCMC approaches for modelers with limited knowledge on the approaches.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Nos. 91225301, 91125021 and 41371473) and China's National Science & Technology Pillar Program (No. 2012BAC03B02). We also thank Dr. Jasper A. Vrugt at University of California, Irvine for providing us the source code of DREAM$_{(ZS)}$.

Appendix A

The joint posterior probability distribution function (pdf) of the model parameters $\boldsymbol{\theta}$ and hyper-parameters $\boldsymbol{\varphi}$ can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z}) \tag{A.1}$$

where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\varphi})$ are the priors and $L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})$ denotes the joint likelihood of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. The marginal posterior pdf $p(\boldsymbol{\theta}|\mathbf{Z})$ equals to the integration of the joint posterior pdf $p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z})$ over $\boldsymbol{\varphi}$, and therefore we have

$$p(\boldsymbol{\theta}|\mathbf{Z}) = \int p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z})d\boldsymbol{\varphi} \propto p(\boldsymbol{\theta}) \int p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})d\boldsymbol{\varphi} \tag{A.2}$$

If a complicated error model (like the one considered in this study) is involved, it may be very difficult to calculate the above integral analytically, and the integral may have to be approximated numerically.

For large-sized sample problems, according to Jeffreys [57], Zellner [58] and Chow [59], the posterior density $p(\boldsymbol{\varphi}\boldsymbol{\theta}, \mathbf{Z})$ can be defined as a Taylor series expansion around the maximum likelihood estimate $\hat{\boldsymbol{\varphi}}$

$$\begin{aligned} p(\boldsymbol{\varphi}|\boldsymbol{\theta}, \mathbf{Z}) &= \frac{p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})}{\int p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})d\boldsymbol{\varphi}} \\ &= (2\pi)^{-\frac{m}{2}}|\mathbf{V}|^{\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}})'\mathbf{V}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}})}\left[1 + o\left(n^{-\frac{1}{2}}\right)\right] \end{aligned} \tag{A.3}$$

where m is the dimension of $\boldsymbol{\varphi}$; n is the number of observations; $\mathbf{V}$ is the inverse covariance matrix and its elements are $\mathbf{V}_{ij} = \frac{\partial^2 \log L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})}{\partial\varphi_i\partial\varphi_j}|_{\hat{\boldsymbol{\varphi}}}$; and $|\mathbf{V}|$ denotes the determinant of $\mathbf{V}$. Note that, for a given realization of $\boldsymbol{\theta}$, its corresponding $\hat{\boldsymbol{\varphi}}$ is at the center of the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})$. Thus, the distribution of $\hat{\boldsymbol{\varphi}}$ should have the same central tendency as the posterior distribution of $\boldsymbol{\varphi}$. Evaluating (A.3) at $\boldsymbol{\varphi} = \hat{\boldsymbol{\varphi}}$ and taking natural logarithms, we have

$$\begin{aligned} \log \int p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})d\boldsymbol{\varphi} &= \log L(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}, \mathbf{Z}) + \log p(\hat{\boldsymbol{\varphi}}) \\ &+ \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}| - \frac{m}{2}\log n + o\left(n^{-\frac{1}{2}}\right) \end{aligned} \tag{A.4}$$

where $\mathbf{R} = \mathbf{V}/n$. $\mathbf{R}$ tends to be a constant matrix when n is a large number. Therefore, the third, fourth, and fifth terms on the right side of Equation A.4 are approximately constant when n is very large, in which case the integral on the left can be approximated as follows

$$\log \int p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})d\boldsymbol{\varphi} \approx \log L(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}, \mathbf{Z}) + \log p(\hat{\boldsymbol{\varphi}}) + C \tag{A.5}$$

where C denotes a constant scalar. Equation A.5 can also be written as

$$\int p(\boldsymbol{\varphi})L(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mathbf{Z})d\boldsymbol{\varphi} \propto L(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}, \mathbf{Z})p(\hat{\boldsymbol{\varphi}}) \tag{A.6}$$

With Eqs. A.2 and A.6, we can have

$$p(\boldsymbol{\theta}|\mathbf{Z}) \propto p(\boldsymbol{\theta})p(\hat{\boldsymbol{\varphi}})L(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}, \mathbf{Z}) = p(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}|\mathbf{Z}) \tag{A.7}$$

Thus, it is clearly that the mathematical nature of the MLE-based inference performed in this study is to sample realizations of $\boldsymbol{\theta}$ from the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{Z})$, which is approximated by $p(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}|\mathbf{Z})$, instead of the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{Z})$.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2015.12.007.

References

[1] Arnold JG, Srinivasan R, Muttiah RS, Williams JR. Large area hydrologic modeling and assessment part I: model development. J Amer Water Resour Assoc 1998;34:73–89. http://dx.doi.org/10.1111/j.1752-1688.1998.tb05961.x.

[2] Neitsch SL, Arnold JG, Kiniry JR, Williams JR. Soil and water assessment tool theoretical documentation version 2009. Texas Water Resources Institute2011.

[3] Chen C, Herr J, Weintraub L. Decision support system for stakeholder involvement. J Environ Eng 2004;130:714–21. http://dx.doi.org/10.1061/(ASCE)0733-9372(2004)130:6(714).

[4] Keller AA, Zheng Y, Robinson TH. Determining critical water quality conditions for inorganic nitrogen in dry, semi-urbanized watersheds. J Amer Water Resour Assoc 2004;40:721–35. http://dx.doi.org/10.1111/j.1752-1688.2004.tb04455.x.

[5] Zheng Y, Keller AA. Stochastic watershed water quality simulation for TMDL development - a case study in the Newport Bay watershed. J Amer Water Resour Assoc 2008;44:1397–410.

[6] Bicknell BR, Imhoff JC, Kittle JL, Jobes TH, Donigian AS. Hydrological simulation program - FORTRAN (HSPF) version 12, user's manual. U.S. Environmental Protection Agency; 2001.

[7] Fonseca A, Botelho C, Boaventura RA, Vilar VJ. Integrated hydrological and water quality model for river management: a case study on Lena River. The Sci Total Environ 2014;485-486:474–89. http://dx.doi.org/10.1016/j.scitotenv.2014.03.111.

[8] Kang MS, Park SW, Lee JJ, Yoo KH. Applying SWAT for TMDL programs to a small watershed containing rice paddy fields. Agric Water Manag 2006;79:72–92. http://dx.doi.org/10.1016/j.agwat.2005.02.015.

[9] Parajuli PB, Mankin KR, Barnes PL. Applicability of targeting vegetative filter strips to abate fecal bacteria and sediment yield using SWAT. Agric Water Manag 2008;95:1189–200. http://dx.doi.org/10.1016/j.agwat.2008.05.006.

[10] Zhang P, Liu Y, Pan Y, Yu Z. Land use pattern optimization based on CLUE-S and SWAT models for agricultural non-point source pollution control. Math Comput Model 2013;58:588–95. http://dx.doi.org/10.1016/j.mcm.2011.10.061.

[11] Zheng Y, Keller AA. Uncertainty assessment in watershed-scale water quality modeling and management: 2. Management objectives constrained analysis of uncertainty (MOCAU). Water Resourc Res 2007;43. http://dx.doi.org/10.1029/2006wr005346.

[12] Beck MB. Water quality modeling: a review of the analysis of uncertainty. Water Resourc Res 1987;23:1393–442. http://dx.doi.org/10.1029/WR023i008p01393.

[13] Griensven Av, Meixner T. Methods to quantify and identify the sources of uncertainty for river basin water quality models. Water Sci Technol 2006;53:51. http://dx.doi.org/10.2166/wst.2006.007.

[14] Rode M, Arhonditsis G, Balin D, Kebede T, Krysanova V, van Griensven A, et al. New challenges in integrated water quality modelling. Hydrol Proc 2010;24:3447–61. http://dx.doi.org/10.1002/hyp.7766.

[15] van Straten G. Models for water quality management: the problem of structural change. Water Sci Technol 1998;37:103–11. http://dx.doi.org/10.1016/s0273-1223(98)00061-4.

[16] Zheng Y, Keller AA. Uncertainty assessment in watershed-scale water quality modeling and management: 1. Framework and application of generalized likelihood uncertainty estimation (GLUE) approach. Water Resources Res 2007;43. http://dx.doi.org/10.1029/2006wr005345.

[17] Montanari A, Shoemaker CA, van de Giesen N. Introduction to special section on uncertainty assessment in surface and subsurface hydrology: An overview of issues and challenges. Water Resourc Res 2009;45 doi: Artn W00b00 Doi 10.1029/2009wr008471.

[18] Stow CA, Reckhow KH, Qian SS, Lamon EC, Arhonditsis GB, Borsuk ME, et al. Approaches to evaluate water quality model parameter uncertainty for adaptive TMDL implementation. J Amer Water Resour Assoc 2007;43:1499–507. http://dx.doi.org/10.1111/j.1752-1688.2007.00123.x.

[19] Honti M, Stamm C, Reichert P. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resources Res 2013;49:4866–84. http://dx.doi.org/10.1002/wrcr.20374.

[20] Kuczera G, Parent E. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. J Hydrol 1998;211:69–85. http://dx.doi.org/10.1016/s0022-1694(98)00198-x.

[21] Schaefli B, Talamba DB, Musy A. Quantifying hydrological modeling errors through a mixture of normal distributions. J Hydrol 2007;332:303–15. http://dx.doi.org/10.1016/j.jhydrol.2006.07.005.

[22] Schoups G, Vrugt JA. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour Res 2010;46. http://dx.doi.org/10.1029/2009wr008933.

[23] Shafii M, Tolson B, Matott LS. Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study. Stochastic Environ Res Risk Assess 2014;28:1493–510. http://dx.doi.org/10.1007/s00477-014-0855-x.

[24] Smith TJ, Marshall LA. Bayesian methods in hydrologic modeling: a study of recent advancements in Markov chain Monte Carlo techniques. Water Resour Res 2008;44. http://dx.doi.org/10.1029/2007wr006705.

[25] Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour Res 2008;44. http://dx.doi.org/10.1029/2007wr006720.

[26] Yang J, Reichert P, Abbaspour KC, Yang H. Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. J Hydrol 2007;340:167–82. http://dx.doi.org/10.1016/j.jhydrol.2007.04.006.

[27] Gardner KK, McGlynn BL, Marshall LA. Quantifying watershed sensitivity to spatially variable N loading and the relative importance of watershed N retention mechanisms. Water Resour Res 2011;47. http://dx.doi.org/10.1029/2010wr009738.

[28] McIntyre N, Jackson B, Wade AJ, Butterfield D, Wheater HS. Sensitivity analysis of a catchment-scale nitrogen model. J Hydrol 2005;315:71–92. http://dx.doi.org/10.1016/j.jhydrol.2005.04.010.

[29] Raat KJ, Vrugt JA, Bouten W, Tietema A. Towards reduced uncertainty in catchment nitrogen modelling: quantifying the effect of field observation uncertainty on model calibration. Hydrol Earth Syst Sci 2004;8:751–63. http://dx.doi.org/10.5194/hess-8-751-2004.

[30] Wellen C, Arhonditsis GB, Long T, Boyd D. Quantifying the uncertainty of nonpoint source attribution in distributed water quality models: a Bayesian assessment of SWAT's sediment export predictions. J Hydrol 2014;519:3353–68. http://dx.doi.org/10.1016/j.jhydrol.2014.10.007.

[31] Balin Talamba D, Parent E, Musy A. Bayesian multiresponse calibration of TOPMODEL: application to the Haute–Mentue catchment, Switzerland. Water Resour Res 2010;46. http://dx.doi.org/10.1029/2007wr006449.

[32] Micevski T, Lerat J, Kavetski D, Thyer M, Kuczera G. Exploring the utility of multiresponse calibration in river system modelling. In: Proceedings of the 19th International Congress on Modelling and Simulation; 2011. p. 3889–95.

[33] Laloy E, Vrugt JA. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing. Water Resour Res 2012;48. http://dx.doi.org/10.1029/2011wr010608.

[34] Vrugt JA, ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D. Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int J Nonlin Sci Numer Simul 2009;10:273–90. http://dx.doi.org/10.1515/ijnsns.2009.10.3.273.

[35] Beven KJ, Smith PJ, Freer JE. So just why would a modeller choose to be incoherent? J Hydrol 2008;354:15–32. http://dx.doi.org/10.1016/j.jhydrol.2008.02.007.

[36] Bates BC, Campbell EP. A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour Res 2001;37:937–47. http://dx.doi.org/10.1029/2000wr900363.

[37] Evin G, Kavetski D, Thyer M, Kuczera G. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resour Res 2013;49:4518–24. http://dx.doi.org/10.1002/wrcr.20284.

[38] Smith T, Sharma A, Marshall L, Mehrotra R, Sisson S. Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. Water Resour Res 2010;46. http://dx.doi.org/10.1029/2010wr009514.

[39] USEPA. Total Maximum Daily Loads for Nutrients San Diego Creek and Newport Bay, California. 1998.

[40] Akhavan S, Abedi-Koupai J, Mousavi S-F, Afyuni M, Eslamian S-S, Abbaspour KC. Application of SWAT model to investigate nitrate leaching in Hamadan–Bahar Watershed, Iran. Agric Ecosyst Environ 2010;139:675–88. http://dx.doi.org/10.1016/j.agee.2010.10.015.

[41] Lam QD, Schmalz B, Fohrer N. Modelling point and diffuse source pollution of nitrate in a rural lowland catchment using the SWAT model. Agric Water Manag 2010;97:317–25. http://dx.doi.org/10.1016/j.agwat.2009.10.004.

[42] Laurent F, Ruelland D. Assessing impacts of alternative land use and agricultural practices on nitrate pollution at the catchment scale. J Hydrol 2011;409:440–50. http://dx.doi.org/10.1016/j.jhydrol.2011.08.041.

[43] Nossent J, Bauwens W. Multi-variable sensitivity and identifiability analysis for a complex environmental model in view of integrated water quantity and water quality modeling. Water Sci Technol: A J Int Assoc Water Pollut Res 2012;65:539–49. http://dx.doi.org/10.2166/wst.2012.884.

[44] Pisinaras V, Petalas C, Gikas GD, Gemitzi A, Tsihrintzis VA. Hydrological and water quality modeling in a medium-sized basin using the Soil and Water Assessment Tool (SWAT). Desalination 2010;250:274–86. http://dx.doi.org/10.1016/j.desal.2009.09.044.

[45] Arnold JG, Kiniry JR, Srinivasan R, Williams JR, Haney EB, Neitsch SL. Soil and water assessment tool Input/Output file documentation version 2009. Texas Water Resources Institute; 2011.

[46] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. Environ Model Softw 2007;22:1509–18. http://dx.doi.org/10.1016/j.envsoft.2006.10.004.

[47] Morris MD. Factorial sampling plans for preliminary computational experiments. Technometrics 1991;33:161–74. http://dx.doi.org/10.2307/1269043.

[48] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I — a discussion of principles. J Hydrol 1970;10:282–90. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

[49] ter Braak CJF, Vrugt JA. Differential evolution Markov chain with snooker updater and fewer chains. Stat Comput 2008;18:435–46. http://dx.doi.org/10.1007/s11222-008-9104-9.

[50] Hantush MM, Chaudhary A. Bayesian framework for water quality model uncertainty estimation and risk management. J Hydrol Eng 2014;19:04014015. http://dx.doi.org/10.1061/(asce)he.1943-5584.0000900.

[51] Duan Q, Sorooshian S, Gupta VK. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J Hydrol 1994;158:265–84. http://dx.doi.org/10.1016/0022-1694(94)90057-4.

[52] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457–72. http://dx.doi.org/10.1214/ss/1177011136.

[53] Zheng Y, Han F. Markov Chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and management. Stoch Environ Res Risk Assess 2015. http://dx.doi.org/10.1007/s00477-015-1091-8.

[54] Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. Water Resour Res 2009;45. http://dx.doi.org/10.1029/2008wr006825.
[55] Keller AA, Chen X, Fox J, Fulda M, Dorsey R, Seapy B, et al. Attenuation coefficients for water quality trading. Environ Sci Technol 2014;48:6788–94. http://dx.doi.org/10.1021/es500202x.
[56] Wu B, Zheng Y, Wu X, Tian Y, Han F, Liu J, et al. Optimizing water resources management in large river basins with integrated surface water-groundwater modeling: a surrogate-based approach. Water Resour Res 2015;51:2153–73. http://dx.doi.org/10.1002/2014wr016653.
[57] Jeffreys H. Theory of probability. Oxford; 1961. p. 193.
[58] Zellner A. An introduction to Bayesian inference in econometrics. New York: John Wiley and sons; 1971. p. 31–4.
[59] Chow GC. A comparison of the information and posterior probability criteria for model selection. J Econom 1981:16.