Environmental Modelling & Software 105 (2018) 286–295

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

On the practical usefulness of least squares for assessing uncertainty in hydrologic and water quality predictions

D. Del Giudice [a, b, *], R. Logsdon Muenich [c], M. McCahon Kalcic [d], N.S. Bosch [e], D. Scavia [f], A.M. Michalak [a]

[a] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA
[b] Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC 27695, USA
[c] School of Sustainable Engineering & the Built Environment, Arizona State University, Tempe, AZ 85287, USA
[d] Department of Food, Agricultural and Biological Engineering, The Ohio State University, Columbus, OH 43210, USA
[e] Lilly Center for Lakes & Streams, Grace College, Winona Lake, IN 46590, USA
[f] University of Michigan, Ann Arbor, MI 48104, USA

ARTICLE INFO

Article history:
Received 17 January 2018
Received in revised form
28 February 2018
Accepted 15 March 2018

Keywords:
Uncertainty assessment
Mechanistic modeling
Surface hydrology
Water quality
Least squares
Statistical inference

ABSTRACT

Sophisticated methods for uncertainty quantification have been proposed for overcoming the pitfalls of simple statistical inference in hydrology. The implementation of such methods is conceptually and computationally challenging, however, especially for large-scale models. Here, we explore whether there are circumstances in which simple approaches, such as least squares, produce comparably accurate and reliable predictions. We do so using three case studies, with two involving a small sewer catchment with limited calibration data, and one an agricultural river basin with rich calibration data. We also review additional published case studies. We find that least squares performs similarly to more sophisticated approaches such as a Bayesian autoregressive error model in terms of both accuracy and reliability if calibration periods are long or if the input data and the model have minimal bias. Overall, we find that, when mindfully applied, simple statistical methods such as LS can still be useful for uncertainty quantification.

1. Introduction

Statistical model calibration and uncertainty quantification (UQ) have recently received substantial attention in surface hydrology and water quality research. Several studies have stressed the importance of more realistically describing the behavior of calibration errors, a.k.a. residuals, and thus moving beyond least squares (LS) calibration assumption of independent and normally distributed residuals (e.g., Reichert and Mieleitner, 2009; Renard et al., 2011; Honti et al., 2013). In particular, it has been suggested that, by using error models that explicitly consider the heteroscedasticity (i.e. non-constant variance) and autocorrelation of the calibration errors, parameter estimation and subsequent predictive uncertainty assessment can be improved in a relatively straightforward manner (Sorooshian and Dracup, 1980; Yang et al., 2007b; Schoups and Vrugt, 2010; Del Giudice et al., 2015a). However, while such sophisticated approaches have been shown to be helpful in the specific situations where they were tested, this does not necessarily imply that simpler statistical techniques such as least squares calibration are never useful. Therefore, there is a critical knowledge gap in hydrologic and environmental modeling regarding when simple calibration approaches are acceptable versus when more sophisticated ones are needed. Understanding the domain of applicability of simple least squares along with its limitations is essential. Indeed, we argue that the presupposition that statistical inference always has to be conducted with conceptually and computationally burdensome methods might be inducing modelers to eschew UQ altogether (e.g., Bosch et al., 2011; Coutu et al., 2012; Razavi and Tolson, 2013) or use "pseudo-statistical" methods with unclear probabilistic interpretation (e.g., Freni et al., 2009; Beven and Smith, 2015). In the context of predictive UQ, we therefore address an important yet so far unanswered question: Are there cases in which a simple method such as least squares yields predictions with precision and accuracy that are on par with state-of-the-art approaches that account for error autocorrelation

* Corresponding author. Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA.
E-mail address: ddelgiudice@carnegiescience.edu (D. Del Giudice).

https://doi.org/10.1016/j.envsoft.2018.03.009
1364-8152/

and parameter uncertainty? If so, are there features that make a particular case study a better candidate for simple approaches to uncertainty quantification? Here, we contrast three case studies with different degrees of data availability and model discrepancy to answer these questions. By also drawing on published hydrologic and water quality case studies we argue that LS can be used, provided that some criteria are met and that the method is applied with some caution. We additionally shed light upon the specific cases in which more sophisticated statistical methods are needed to deliver useful parameter estimates and uncertainty intervals.

2. Study areas and models

We investigate the suitability of LS calibration and subsequent uncertainty propagation in two catchments that differ substantially in terms of geographic domain, availability of calibration data, quality of input data, type and complexity of hydrological model, variables predicted, and types of systematic errors. For one catchment we consider two cases, one with low systematic deviations between model results and output data and one with high bias induced by forcing the model with less accurate input estimates. In each of the three cases we split the recorded time series into a calibration period, where output data are used for parameter estimation, and a validation period, where output data are used to corroborate model predictive abilities.

2.1. Case studies 1 (CS1 and CS1'): watershed with limited data

CS1 and CS1′ involve the same small, partially combined sewer network located in Adliswil, Zurich Canton, Switzerland (Fig. S1). The watershed has an area of 28.6 ha, only a fraction of which contributes to the sewer outflow. The effective contributing area of the watershed is indeed a calibration parameter (see below). The area is characterized by medium density residential development and a slope of about 8.7%. The site was monitored in 2013 to quantify the occurrence of sewer overflows and to understand the impact of the location of precipitation measurements on discharge predictions. For calibration we use a discharge Q [l/s] time series of an event including 97 observations recorded every 4 min (Fig. 1). This calibration period includes two storm events of duration greater than the catchment response time, which is on the order of minutes. For validation we use a subsequent event that occurred 80 days later and included 179 observations. Such short time series are typical in hydrological modeling of urban catchments (e.g., Freni et al., 2009; Coutu et al., 2012). For CS1, input data were recorded by a pluviometer from the Swiss meteorological office[1] located circa 7.5 km Northeast of the catchment (Fig. S1). The second version of this case study, CS1', uses more accurate input obtained by averaging data from two pluviometers located within the catchment area itself. A comparison of the precipitation records from these pluviometers reveals that the precipitation input data used in CS1 has substantial systematic errors (Fig. S3). The time series of sewer runoff at the outlet of the catchment is modeled using a lumped linear reservoir model with a harmonic function describing the wastewater oscillations (see Del Giudice et al. (2016) for further details about the catchment and the model). In this investigation, we calibrate the three model parameters related to rainfall-runoff, namely A [m^2], the area contributing to the stormwater outflow, k [hr], the mean residence time in the virtual reservoir representing the catchment, and x_{gw} [l/s], the baseflow.

[1] www.hw.zh.ch/hochwasser/foto/DB%20SMA.pdf.

2.2. Case study 2 (CS2): watershed with abundant data

CS2 is the River Raisin basin, which has an area of 2784 km^2 and is primarily rural (72%) and forested (16%). The variables of interest are river discharge Q [m^3/s] and soluble reactive phosphorus load *SRP*, [kg/d]. The calibration period contains 1095 discharge observations and 1095 *SRP* load observations at daily resolution (Fig. 3). This calibration period includes numerous storm events of duration longer than the catchment response time, which is on the order of days. The validation period immediately follows the calibration period and includes 366 discharge observations and 335 *SRP* load observations. The watershed dynamics are simulated using the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998). SWAT is a hydrologic transport model that operates at catchment scale. It is both more complex, due to its more explicit representation of spatial heterogeneity and watershed processes, and more computationally-demanding than the simple reservoir model used in CS1 and CS1'. The River Raisin basin and model are well studied in the context of furthering the understanding of the dynamics of nutrient loading from agricultural areas (Bosch et al., 2011). The SWAT model used here includes all the same process parameterizations, inputs, and management details as in Muenich et al. (2017). The model is driven by daily precipitation and temperature observations from nine NOAA GHCN land surface stations (Menne et al., 2012), most of which are located within the catchment area (Fig. S2). Daily discharge and *SRP* observations used for calibration and validation are obtained from Heidelberg University NCWQR (2015). In the current application, we calibrate three model parameters: *CN*2 [-], the runoff curve number for moisture condition II, *SMTMP* [°C], the snow melt base temperature, and *PHOSKD* [-], the phosphorus soil partitioning coefficient. These parameters are selected because they are primary controls on three key processes, namely rainfall-runoff, snowmelt, and biochemical reaction, and the output variables of interest are sensitive to them.

3. Methods

3.1. Simple method: frequentist least squares (LS)

The least squares method, LS, is a classic statistical approach for calibrating model parameters, estimating model output errors, and thus producing prediction intervals (Wooldridge, 2015). LS is generally adopted as the basic technique against which new methods for uncertainty quantification are tested (Sorooshian and Dracup, 1980; Schoups and Vrugt, 2010; Renard et al., 2011; Honti et al., 2013; Del Giudice et al., 2016). The simplest application of LS is within a frequentist framework, in which model parameters are assumed to have one true yet unknown value. Consequently, model parameters are estimated by minimizing an objective function and neither prior nor posterior model parameter uncertainties are explicitly considered. Because model residuals in hydrology are typically heteroskedastic and non-normal (Wang et al., 2012; Del Giudice et al., 2013), here we apply LS after having transformed the observed $\mathbf{y}_o$ and modeled $\mathbf{y}$ output using a non-linear monotonic function g (see Supporting Material). The objective function used for calibration is the sum of the squares of the errors:

$$SSE = \sum^{n} (\tilde{\mathbf{y}}_o - \tilde{\mathbf{y}})^2 \quad (1)$$

where tilde represents the transformed output and n is the number of data points in the calibration dataset, i.e. the length of $\mathbf{y}_o$, a vector possibly including multiple outputs. Numerically, we use an adaptive Markov chain Monte Carlo algorithm (as in Del Giudice

Table 1
Conceptual model and error model calibration parameters (θ). The notation for prior distributions is: LN (μ,σ): lognormal, N (μ,σ): normal, TN (μ,σ,a_1,a_2): truncated normal, Exp (λ^{-1}): exponential. The symbols are: μ: expected value, σ: standard deviation, a_1: lower limit, a_2: upper limit, λ: rate. g represents the output transformation function and $\frac{dg}{dy}\Big|_{y^*}$ the derivative of the transformation evaluated in y^* (see SI).

Variable	Description	Units	Prior
CS1: linear reservoir model			
A	area contributing to outflow	[m^2]	LN (11815.8, 1181.6)
k	water residence time	[hr]	LN (0.079, 0.016)
x_{gw}	groundwater infiltration and sewage baseflow	[l/s]	LN (2.05, 0.013)
σ_{E_Q}	σ of the uncorrelated outflow errors	[g (l/s)]	$\mathrm{LN}\left(4.1\frac{dg}{dy}\Big\|_{50},\ 0.41\frac{dg}{dy}\Big\|_{50}\right)$
σ_{B_Q}	σ of the autocorrelated outflow errors	[g (l/s)]	$\mathrm{TN}\left(0, 3.77\frac{dg}{dy}\Big\|_{50},\ 0, \infty\right)$
τ	correlation length of the outflow errors	[hr]	LN (0.47, 0.047)
CS2: SWAT model			
CN2	runoff curve number for moisture condition II	[-]	TN (-0.04, 0.15, -0.25, 0.15)
SMTMP	snow melt base temperature	[°C]	TN (-2.1, 2, -5, 5)
PHOSKD	phosphorus soil partitioning coefficient	[-]	TN (160, 10, 100, 175)
σ_{E_Q}	σ of the uncorrelated outflow errors	[g (m^3/s)]	$\mathrm{LN}\left(10^{-1}\frac{dg}{dy}\Big\|_{100},\ 5\cdot10^{-2}\frac{dg}{dy}\Big\|_{100}\right)$
σ_{B_Q}	σ of the uncorrelated outflow errors	[g (m^3/s)]	$\mathrm{TN}\left(0, 10\frac{dg}{dy}\Big\|_{100},\ 0, \infty\right)$
σ_{E_P}	σ of the uncorrelated *SRP* loading errors	[g (kg/d)]	$\mathrm{LN}\left(1.8\frac{dg}{dy}\Big\|_{900},\ 0.9\frac{dg}{dy}\Big\|_{900}\right)$
σ_{B_P}	σ of the autocorrelated *SRP* loading errors	[g (kg/d)]	$\mathrm{TN}\left(0, 180\frac{dg}{dy}\Big\|_{900}, 0, \infty\right)$
τ	correlation length of the output errors	[d]	LN (7, 3.5)

et al. (2013)) to find the parameter set that minimizes *SSE*. After model calibration, model predictions for the validation period are obtained by running the model with the best estimates of the parameters, and 95% uncertainty intervals (a.k.a. prediction intervals for new observations) are approximated as plus/minus two times the root mean squared error observed during the calibration period (Wooldridge, 2015). The upper and lower 95% uncertainty intervals for each output k, here being Q or *SRP*, are then back-transformed using a function g^{-1} to obtain the 95% interquantile range (*IQR*) in the original space:

$$IQR_{95,k} \approx g^{-1}\left(\tilde{\mathbf{y}}_{val,k} \pm 2\sqrt{\frac{1}{n_k}\sum^{n_k}\left(\tilde{\mathbf{y}}_{o,k} - \tilde{\mathbf{y}}_k\right)^2}\right) \quad (2)$$

where $\tilde{\mathbf{y}}_{val,k}$ represents the (transformed) model output of type k in the validation period, and n_k is the number of data points in the calibration dataset for the output of type k.

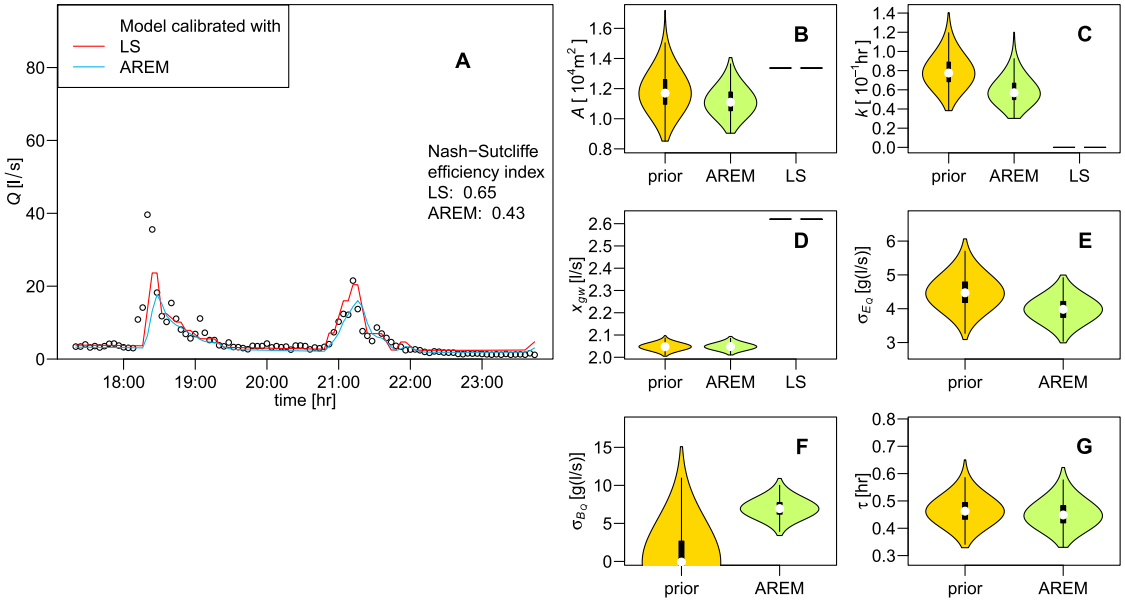

Fig. 1. Calibration results for CS1 obtained with the two methods, LS and AREM. (A) Time series of the model output using the optimized parameters (red line for LS) or the median of the posteriors (light blue line for AREM). Calibration data are represented by dots. (B–D) Model parameters calibrated with the two methods. While AREM makes use of the prior information to produce posterior parameter distributions (both prior and posterior distributions are represented by violin plots), LS generates parameter values which minimize the objective function *SSE*. (E–G) Error model parameters for AREM. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.2. Complex method: Bayesian autoregressive error model (AREM)

As a prototypical example of a more complex statistical technique, we here choose a Bayesian approach that represents model discrepancy or bias as a Gaussian process. We call this method AutoRegressive Error Model (AREM). The key difference between AREM and LS consists in the consideration of output error autocorrelation. AREM derives from applied statistics (Kennedy and O'Hagan, 2001; Bayarri et al., 2007; Reichert and Schuwirth, 2012) and here we adopt an implementation that was recently proposed in the hydrological literature (Del Giudice et al., 2013). The approach considers the autocorrelation in residuals resulting from model structural deficits (linked to oversimplifications in the process description and insufficient spatial discretization (Reichert and Mieleitner, 2009)) and input errors (due to imperfect estimation of time-varying inputs such as precipitation (Del Giudice et al., 2016)). As discussed earlier, the concept of stochastically describing error autocorrelation has helped improve parameter estimation and model predictions in several hydrological contexts (Sorooshian and Dracup, 1980; Yang et al., 2007a; Schoups and Vrugt, 2010; Del Giudice et al., 2015a; Sikorska et al., 2015). Besides its popularity, AREM has the advantage of realistically describing predictive uncertainty even in the presence of model bias, while still being applicable in conjunction with computationally-expensive aquatic models (Del Giudice et al., 2015b; Dietzel and Reichert, 2014).Model calibration involves characterizing the posterior distribution of both hydrological model and error model parameters:

$$f(\boldsymbol{\theta}|\mathbf{y}_o) = \frac{f(\boldsymbol{\theta})f(\mathbf{y}_o|\boldsymbol{\theta})}{\int f(\boldsymbol{\theta})f(\mathbf{y}_o|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \quad . \tag{3}$$

where $f(\boldsymbol{\theta})$ represents the prior parameter distribution and $f(\mathbf{y}_o|\boldsymbol{\theta})$ is the likelihood function that represents the errors as a sum of white noise and a Gauss-Markov process. The Gauss-Markov process is the continuous time equivalent of an autoregressive process of order one, a parametrization used in other applications of AREM (e.g., Evin et al., 2014). The likelihood function embodies assumptions about the output distribution and, in a Bayesian framework, it enables us to extract information from the data $\mathbf{y}_o$ about parameters $\boldsymbol{\theta}$ of the deterministic model and of the error model. For AREM, the likelihood function $f(\mathbf{y}_o|\boldsymbol{\theta})$ is Gaussian and centered at the deterministic model output $\tilde{\mathbf{y}}(\boldsymbol{\theta})$ in a transformed space (Del Giudice et al., 2015b; Sikorska et al., 2015):

$$f(\mathbf{y}_o|\boldsymbol{\theta}) = \frac{(2\pi)^n}{\sqrt{\det(\boldsymbol{\Sigma}(\boldsymbol{\theta}))}}\exp\Bigg(-\frac{1}{2}[\tilde{\mathbf{y}}_o - \tilde{\mathbf{y}}(\boldsymbol{\theta})]^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}[\tilde{\mathbf{y}}_o - \tilde{\mathbf{y}}(\boldsymbol{\theta})] \times \Bigg)\prod_{i=1}^{n}\frac{\mathrm{dg}}{\mathrm{dy}}\left(y_{o,i}\right) \tag{4}$$

where $g(\,)$ represents the transformation of the output of length n. While in CS1 there is only one output, Q, in CS2 there are two outputs, Q and SRP. In the latter case the joint likelihood function is given by:

$$f\left(\mathbf{y}_o^Q, \mathbf{y}_o^{SRP}\middle|\boldsymbol{\theta}\right) = f\left(\mathbf{y}_o^Q\middle|\boldsymbol{\theta}\right)\cdot f\left(\mathbf{y}_o^{SRP}\middle|\boldsymbol{\theta}\right) \tag{5}$$

The covariance $\boldsymbol{\Sigma}$ is a square matrix of order n:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{i,j} = \sigma_B^2\exp\left(-\frac{|t_i - t_j|}{\tau}\right) + \delta_{ij}\sigma_E^2 \tag{6}$$

where σ_B^2 and σ_E^2 are parameters representing the variance of the Markov bias process and of the white noise process, respectively, τ is the correlation length of the Markov bias process, i and j are subscripts spanning over the calibration time domain, and δ represents the Kronecker delta. While the Markov bias process describes the autocorrelated output errors deriving form a combination of time-dependent input errors and model structural deficits, the white noise process describes the uncorrelated output errors. Note that in CS2 each output variable has a different σ_E and σ_B (Table 1). Viewed from a Bayesian perspective, the calibration framework in Sec. 3.1 is equivalent to maximizing (4) with non-informative priors $f(\boldsymbol{\theta}) \propto 1$ and diagonal covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{i,j} = \delta_{ij}\sigma_E^2$ (Borsuk et al., 2002). Therefore, model calibration is conducted iteratively using the same adaptive Markov chain Monte Carlo algorithm as in Section 3.1, with the difference that here the full representation of the posterior distribution is of interest, rather than just its mode. Model predictions are obtained by propagating a large sample from the posterior parameter distribution through the hydrological model and the error model (Eqs. (26) and (27) in Del Giudice et al. (2016)). The 95% uncertainty intervals for new observations are derived by empirically calculating the 2.5th and 97.5th quantiles at each time point and then transforming those back to the real space via the function g^{-1}.

3.3. Prior distribution of hydrological model and error model parameters

As a Bayesian method, AREM can accommodate prior knowledge about hydrological model and error model parameters. This prior information is typically based on experience with similar models and datasets as those being investigated. Following Del Giudice et al. (2016), for CS1 and CS1' we define the priors of model parameters as lognormal distributions. For CS2, we instead select normal distributions centered at default SWAT values and truncated at values suggested in previous investigations (Muenich et al., 2017). We use lognormal prior distributions for the error model parameters σ_E and τ and a truncated normal distribution for σ_B as suggested by Del Giudice et al. (2015b). The latter has zero both as the mean before truncation and the lower bound, indicating the preference for having most of the variability in the data explained by the model rather than by the bias process. More details on prior parameters are given in Table 1 and prior distributions are shown in Figs. 1 and 3.

4. Results

4.1. Model calibration

For CS1 we observe that the model fits the calibration data substantially better when parameters are optimized using LS rather than AREM (Fig. 1). This is evident by observing the red line, which is closer to the peak discharge observations, and the higher NS, which indicates the ability of the LS-calibrated model to match high value output data more closely. The difference in predictive performances between the methods is attributable primarily to differences in the calibrated parameter estimates between LS and AREM. For CS1 a substantial difference is evident between the parameter estimates obtained with the two methods (Fig. 1). AREM, which incorporates existing knowledge on model parameters, yields posterior distributions similar to the priors. The only parameter that is substantially informed by the calibration process is σ_B, which represents the amount of bias identified. LS parameter estimates, which are not informed by priors, are instead dramatically different from both the priors and the AREM estimates. Interestingly, LS and AREM perform almost identically for the modified version of the case study, CS1′, wherein the precipitation

input data with systematic errors is replaced with more accurate data. The model in CS1' exhibits low bias as demonstrated by visual inspection, high Nash-Sutcliffe efficiency (NS), relatively small standard deviation of the bias term ($\sigma_{B_Q} < \sigma_{E_Q}$) (Fig. 2), and low residual autocorrelation (Fig. S4).

For CS2, the two approaches yield almost identical model fits, showing higher accuracy for discharge than for *SRP* (Fig. 3). Interestingly, even though LS does not make use of prior information, it infers hydrologic parameters comparable to the prior estimates. Overall, LS involved $\sim 10^3$ model simulations, whereas AREM required $\sim 10^4$ model runs. This represents a substantial difference in the calibration cost between the two methods. The higher computational cost associated with AREM is attributable to the need to obtain a representative sample of the full posterior distribution and to the presence of more error model parameters to estimate.

4.2. Model predictions for the validation period

The predictive performance of each method is assessed by comparing the median and the 95% uncertainty intervals for a validation period with an independent dataset of measured outputs. Besides relying on visual assessment, as in previous studies (e.g., Reichert and Mieleitner, 2009; Sikorska et al., 2015; Del Giudice et al., 2016) we use the Nash-Sutcliffe efficiency index to measure predictive accuracy and the percentage of data falling within the 95% prediction intervals to measure the reliability of the uncertainty bands. Reliability (increasing the closer to 95% the actual coverage of validation data is) and precision (increasing with narrower uncertainty bands) of the 95% prediction intervals are also simultaneously assessed by the negative of the interval skill score (Supporting Material). The better the quality of the predictions, the closer to 0 this statistic is. The value of these metrics for our experiments are given in Table 2.

While in the calibration period we observe that LS produces more accurate predictions than AREM for CS1, the situation is reversed during the validation period. The limited calibration data and high systematic precipitation errors lead LS to overfit observations and converge on erroneous parameter estimates, while AREM minimizes this undesirable effect. Besides being more accurate, predictions with AREM are also more reliable, in that the uncertainty bounds are more representative of the true uncertainty (Fig. 4). When reducing the systematic precipitation errors in CS1 (i.e., CS1′), LS again yields results on par with AREM (Fig. 5). In other words, even with limited calibration data, LS performs well as long as systematic input errors are minimal. This is true both in terms of prediction accuracy and reliability. However, AREM uncertainty intervals account for parameter uncertainty, making them wider during high flows, and thus potentially making model predictions even more reliable, especially in those crucial periods.In CS2 the situation is different from CS1 but similar to CS1'. As for the calibration phase, model performance during validation is comparable for the two methods, with AREM having slightly higher predictive accuracy. Predictive uncertainty is also very similar for the two methods, with LS having slightly more reliable and precise confidence intervals (Fig. 6).

5. Discussion and conclusions

5.1. Factors favoring the application of least squares calibration

The results of the three case studies presented here are consistent with published studies where LS was implemented for prediction during a validation period (Table 3) or only during a calibration period (e.g., Borsuk et al., 2002; Vrugt et al., 2003; Freni et al., 2009; Forrest et al., 2011; Wang et al., 2012; Jiang et al., 2015). These studies span a variety of landscapes, aquatic systems and models, ranging from conceptual lumped models to physically-based semi-distributed models. LS is implemented in a variety of ways across the published studies, for instance by assuming the residuals to have a constant variance (e.g., Reichert and Schuwirth, 2012), a constant variance after variable transformation (e.g., Dotto et al., 2011), or a variance linearly dependent on the output magnitude (e.g., Westra et al., 2014). A consistent assumption, however, is that the output errors during the calibration period are independent in time and normally distributed (Wooldridge, 2015). The case studies examined here, together with those published previously, point to two key characteristics that favor the use of

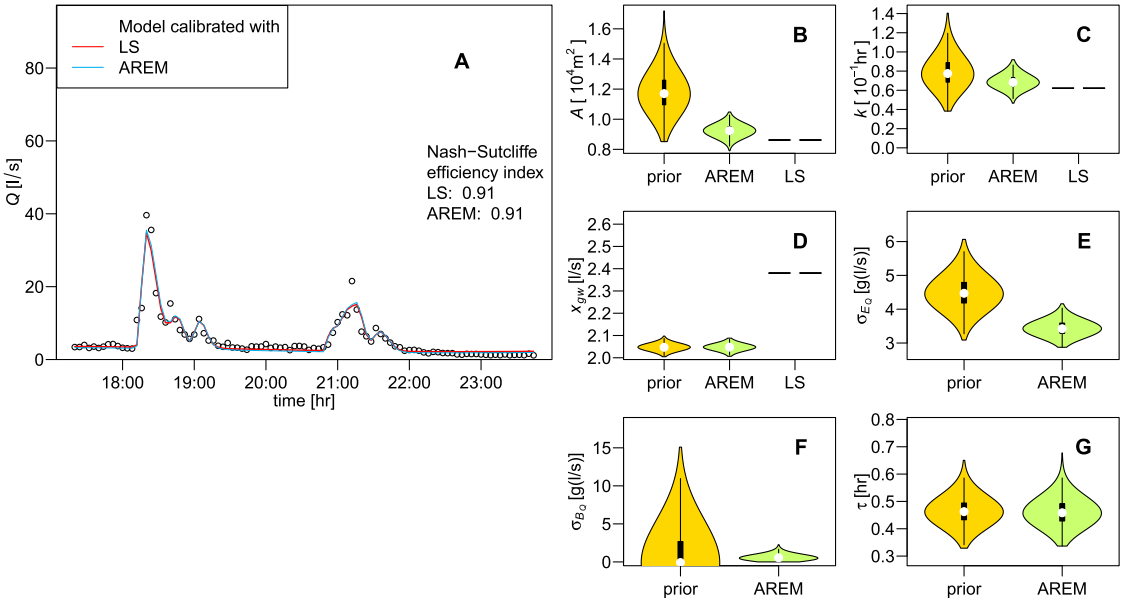

Fig. 2. Calibration results for CS1′ obtained with the two methods, LS and AREM. Same as Fig. 1, yet here the model is forced with accurate precipitation data (Fig. S3A).

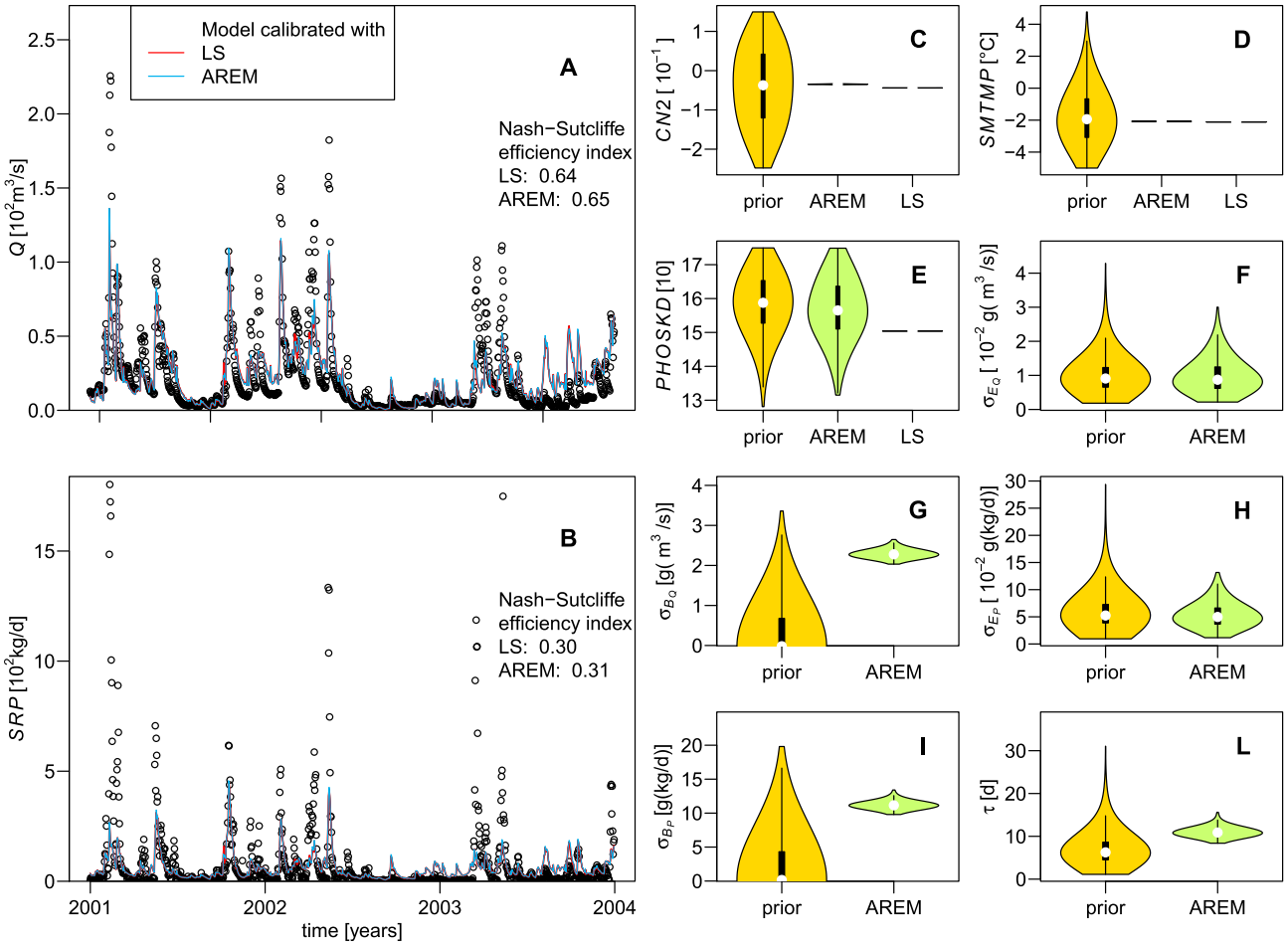

Fig. 3. Calibration results for CS2 obtained with the two methods, LS and AREM. (A, B) Time series of the model outputs using the optimized parameters for LS (red line) or the median of the posteriors for AREM (light blue line). Model outputs obtained with LS and AREM overlap almost perfectly. Calibration data are represented by dots. (C–E) Model parameters calibrated with the two methods. While AREM makes use of the prior information to produce posterior parameter distributions (both prior and posterior distributions are represented by violin plots), LS generates parameter values which minimize the objective function *SSE*. (F–L) Error model parameters for AREM. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Metrics assessing the predictive validation performances for the three case studies (Figs. 4–6). "Coverage" represents the percentage of validation data covered by the 95% uncertainty intervals, "NS" stands for Nash-Sutcliffe efficiency, and "Skill" indicates the negative of the 95% interval skill score (Sect. 4.2). The ideal values for the three metrics are respectively 95 (perfect reliability), 1 (perfect accuracy), and 0 (perfect reliability and precision).

Case	Variable	Method	Coverage	NS	Skill
CS1	*Q*	LS	78%	0.08	−77
		AREM	91%	0.45	−37
CS1′	*Q*	LS	92%	0.84	−20
		AREM	93%	0.84	−19
CS2	*Q*	LS	93%	0.54	−64
		AREM	88%	0.55	−73
	SRP	LS	99%	0.63	−262
		AREM	100%	0.65	−336

simple approaches such as LS. Both of these characteristics are indicative of high information content in the training dataset, making it possible to estimate parameters of both the physical model and the error model in a manner that will yield accurate and reliable predictions (Kavetski et al., 2011; Razavi and Tolson, 2013; Westra et al., 2014; Beven and Smith, 2015).

5.1.1. Long calibration time series

In the case studies using long calibration time series (e.g., CS2, Table 3, Borsuk et al., 2002; Vrugt et al., 2003; Wang et al., 2012; Jiang et al., 2015) model predictions obtained with LS are usually accurate (referring to the median) and reliable (referring to the uncertainty). In the context of watershed modeling, long calibration periods refer to periods that include numerous runoff events and span a variety of hydrologic conditions, typically over the course of multiple months or even years (Razavi and Tolson, 2013). It has previously been shown that LS in conjunction with a long calibration period can produce robust estimates of model parameters and thus accurate predictions (e.g., Yapo et al., 1996; Bosch et al., 2011; Razavi and Tolson, 2013). While these earlier studies had not investigated the ability of LS to produce reliable uncertainty intervals, they did show that long calibration datasets make it more likely that a variety of environmental conditions will be captured by the calibration data, which is critical for calibrating model parameters. Additionally, long calibration time series reduce the impact of any periods that may be less representative, which have also been referred to as "misleading short periods" (Razavi and Tolson, 2013) or "disinformative events" (Beven and Smith, 2015). These considerations are also consistent with Honti et al. (2013) who, having calibrated a hydrological model over a period of almost eight years, found that even in presence of model bias,

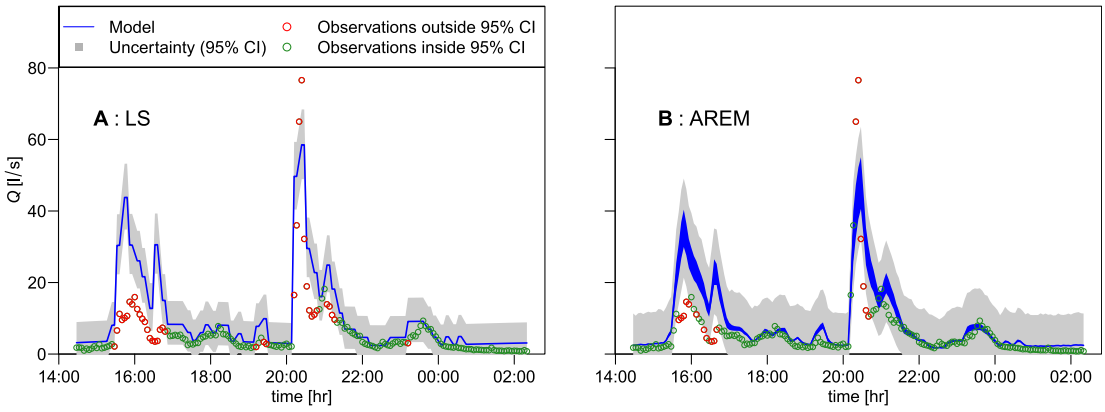

Fig. 4. Discharge predictions for CS1 using both methods, LS (A) and AREM (B). Validation data are shown. The blue area for AREM represents the effect of model parameter uncertainty on predictions (95% confidence intervals, CI). Quantitative assessment of predictions is given in Table 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

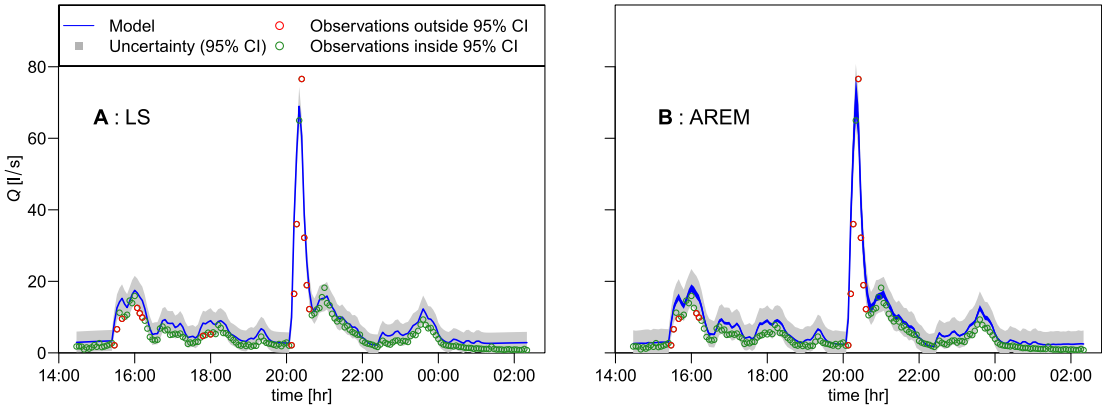

Fig. 5. Discharge predictions for CS1′ using both methods, LS (A) and AREM (B). Same as Fig. 4, yet here the model is forced with accurate precipitation data (Fig. S3B).

predictive accuracy with LS and AREM was equally high. In terms of predictive uncertainty, having longer calibration datasets also makes it possible to better assess the variance of the predictive errors (Kavetski et al., 2011; Wooldridge, 2015). Overall, this finding also highlights the importance of long-term monitoring programs for improved hydrologic predictions, such as the National Center for Water Quality Research (Heidelberg University NCWQR, 2015) from which this study benefited.

5.1.2. Low systematic model errors during calibration period

It is difficult to set a quantitative threshold for low model bias, and several studies therefore mainly discuss the visual observation that model results and calibration data show minimal systematic discrepancies (e.g., Bayarri et al., 2007; Reichert and Schuwirth, 2012). However, in addition to visual inspection, we also consider model bias to be low when NS is close to 1 (Gupta et al., 2009), the identified bias is within the range of the uncorrelated output errors ($\sigma_{B_Q} \leq \sigma_{E_Q}$) (Del Giudice et al., 2016), and residuals show negligible autocorrelation (Yang et al., 2007b). In general, as shown when comparing CS1 and CS1' and the two cases from Reichert and Schuwirth (2012) in Table 3, LS performs better when model bias, due to systematic errors in input data or to structural deficits in the hydrologic/water quality model itself, is low. This is true both in terms of prediction accuracy and reliability. The lack of systematic errors can in fact compensate for short calibration time series. This effect is linked to the fact that, when model bias is negligible, LS is meant to estimate unbiased model parameters even with a small sample size (Wooldridge, 2015) i.e. a short calibration period. However, because surface hydrology and pollutant transport models can exhibit substantial systematic deviations from monitoring data over the short calibration periods available, much attention has been devoted to identifying representative time series that can be used for more robust parameter estimation. For instance, Razavi and Tolson (2013) showed that a short yet representative calibration period can produce LS calibration results as useful for predictions as those produced using long calibration periods. For short representative periods, it is important that the model show low output bias and that the input data include sufficient variability to identify parameters representative of both high and low flow conditions. This is the reason for which, in rainfall-runoff modeling, it is usually recommended that calibration be performed during wet conditions, when the dynamic inputs vary sufficiently to enable appropriate parameter sensitivity and identification (Yapo et al., 1996). Additionally, as discussed by Beven and Smith (2015), it is best to avoid periods where the relationship between input and output data is unusual or inconsistent with typical conditions, as this may be indicative of a period with substantial input errors.

5.1.3. Caveats

Overall, we find that in the presence of long calibration time series, or even with short time series with low systematic input and model errors, LS calibration is likely to lead to accurate and reliable predictions. At the same time we acknowledge that, even when one

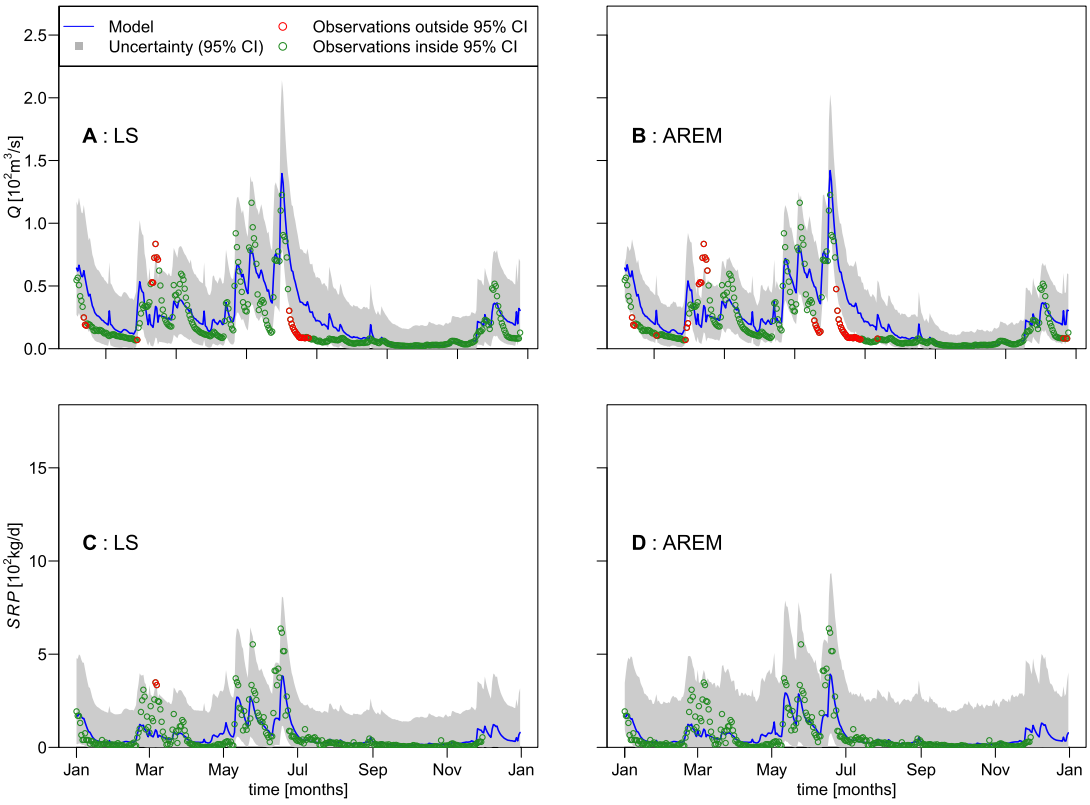

Fig. 6. Discharge (A, B) and nutrient load (C, D) predictions for the River Raisin (CS2) using both methods, LS (A, C) and AREM (B, D). Validation data are shown. Parameter uncertainty for AREM is too small to be visible on panels B and D. Quantitative assessment of predictions is given in Table 2.

Table 3
Hydrologic and water quality case studies having evaluated the predictive performances of LS calibration in an independent validation period. In presence of short calibration periods (italicized terms in the 3rd column) and substantial model biases (italicized terms in the 4th column), predictions in the validation period are unreliable (italicized terms in the 5th column). Instead, outside of these pathological cases, LS predictions are found to be satisfactory. "NS" stands for Nash-Sutcliffe efficiency for the calibration period and "Coverage" represents the percentage of validation data covered by the 95% uncertainty intervals. For the current study "Coverage" is as in Table 2, except for CS2 where the average for both output variables is reported. Note estimates with ≈ are approximate readings from plotted values.

Investigation	Case Study	Calibration Period	Calibration Bias	Coverage
Dotto et al. (2011)	Richmond catchment	multiple events, 2 years	NS=0.71	≈99%
Kavetski et al. (2011)	Weierbach catchment	multiple events, 2 years	NS=0.88	≈95%
Reichert and Schuwirth (2012)	synthetic, unbiased model	*20 points*	visual	≈95%
	synthetic, biased model	*20 points*	*visual*	*≈85%*
Del Giudice et al. (2013)	Sadova catchment	*few events, 2 days*	NS=0.95	89%
Evin et al. (2014)	synthetic, unbiased model	multiple years	discussed	discussed
Westra et al. (2014)	Scott Creek catchment	multiple years	NS=0.80	discussed
Del Giudice et al. (2015a)	Ballerup catchment	*few events, 10 day*	*visual*	*83%*
Current study	CS1	*few hours*	*NS=0.65*	*78%*
	CS1'	*few hours*	NS=0.91	92%
	CS2	multiple events, 3 years	*NS=0.47*	96%

of these conditions is fulfilled, there might be particular situations that require caution. For instance, AREM can provide more reliable predictions than LS in the special case of wanting to quantify the uncertainties at aggregated scales after model calibration has been performed as finer scales (Evin et al., 2014). The reason is that assessing uncertainty at aggregated scales required a quantification of error covariances in addition to variances. Also, AREM can provide more accurate and precise predictions in short term forecasting mode (Del Giudice et al., 2015a) when such periods are on the order of the autocorrelation timescale of the prediction errors, because it considers residual autocorrelation (Evin et al., 2014). Finally, we emphasize the importance of considering error heteroscedasticity by implementing LS calibration as a weighted least squares (WLS) approach (Kavetski et al., 2011; Wooldridge, 2015). While for some aquatic systems the error variance may not depend substantially on the magnitude of the output variable (e.g., Borsuk et al., 2002; Reichert and Schuwirth, 2012; Jiang et al., 2015), most studies focusing on streamflow predictions do report errors that vary with output magnitude (Sorooshian and Dracup, 1980; Honti et al., 2013; Sikorska et al., 2015). Weighted LS is still a very simple approach and can be implemented either by using a data transformation, as done here and elsewhere (Wang et al., 2012; Del Giudice et al., 2013), or a linear heteroscedastic model (Evin et al., 2014; Westra et al., 2014). Several studies underscoring the importance of using more sophisticated error models have actually done so by comparing the performances of ordinary LS against those of approaches that simultaneously account for autocorrelation and heteroscedasticity (e.g., Schoups and Vrugt, 2010; Del

Giudice et al., 2015a). In these cases, it is therefore not clear whether the suboptimal predictions based on LS were simply due to a lack of weighting of the output. Indeed, other studies have discussed how results of ordinary LS can be substantially improved when heteroscedasticity is appropriately accounted for (e.g., Kavetski et al., 2011).

5.2. Summary and outlook

In this study we have analyzed the usefulness of least squares calibration for obtaining accurate and reliable predictions of hydrologic time series. The need for such an analysis has recently arisen in response to investigations suggesting that more sophisticated methods are required for UQ. We have therefore addressed this gap by performing calibration and validation on three case studies and reviewing numerous published studies on the topic. We find that LS produces satisfactory predictions in cases where the calibration period is either long (e.g., includes numerous runoff events) or short but displays low systematic errors, provided that heteroscedasticity is appropriately taken into account. Under these circumstaces, a more complex methods such as AREM provide results that are very similar to LS. At the same time, we have also shown that AREM can be helpful when the model is biased and calibration time series are short. We acknowledge that for particular applications, such as when the goal is to disentangle and reduce the sources of total predictive uncertainty, stochastic methods of even higher complexity than AREM (e.g., Reichert and Mieleitner, 2009; Renard et al., 2011; Del Giudice et al., 2016) may be required. However, if the focus is on accurate and reliable predictions rather than on apportioning total uncertainty, we find that LS can perform well.The guidelines provided here can be particularly beneficial when considered prior to conducting model calibration and uncertainty propagation. Besides providing useful guidance for uncertainty analysis in a statistical framework, we hope our study will foster further research focusing on: i) the formulation of quantitative metrics for defining a priori the information content and representativeness of a calibration period, ii) understanding the role of prior information about model parameters, residual heteroscedasticity, and residual autocorrelation, and iii) further understanding how methods of different complexity perform in the presence of non-stationary errors.

Acknowledgements

The data used are available upon request from ddelgiudice@carnegiescience.edu. This material is based upon work supported by the National Science Foundation under Grant No. (CBET 1313897). Additional funding for Margaret Kalcic was provided by EPA under Grant No. (GL-00E0461-0). We are grateful to Carlo Albert for his thoughts on the initial part of this work, and Wolfgang Nowak, Jasper A. Vrugt, Mary C. Hill, and an anonymous reviewer for their feedback on the manuscript. Discharge and loading data for the River Raisin is available from the Heidelberg University National Center for Water Quality Research. Weather data for the same watershed were obtained from the National Oceanic and Atmospheric Association Global Historical Climatology Network. Data for the Adliswil catchment were from Del Giudice et al. (2016). Precipitation data for CS1 were obtained from MeteoSwiss.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.envsoft.2018.03.009.

References

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part i: model development. JAWRA J. Am. Water Resour. Assoc. 34 (1), 73–89.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., Tu, J., 2007. A framework for validation of computer models. Technometrics 49 (2), 138–154.

Beven, K., Smith, P., 2015. Concepts of information content and likelihood in parameter calibration for hydrological simulation models. J. Hydrol. Eng. 20 (1), A4014010.

Borsuk, M.E., Stow, C.A., Reckhow, K.H., 2002. Predicting the frequency of water quality standard violations: a probabilistic approach for tmdl development, 36 (10), 21092115.

Bosch, N.S., Allan, J.D., Dolan, D.M., Han, H., Richards, R.P., 2011. Application of the soil and water assessment tool for six watersheds of lake erie: model parameterization and calibration. J. Great Lake. Res. 37 (2), 263–271.

Coutu, S., Del Giudice, D., Rossi, L., Barry, D., 2012. Parsimonious hydrological modeling of urban sewer and river catchments. J. Hydrol. 464–465 (0), 477–484.

Del Giudice, D., Albert, C., Rieckermann, J., Reichert, P., 2016. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. Water Resour. Res. 52 (4), 3162–3186.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., Rieckermann, J., 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. Hydrol. Earth Syst. Sci. 17 (10), 4209–4225.

Del Giudice, D., Löwe, R., Madsen, H., Mikkelsen, P.S., Rieckermann, J., 2015a. Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors. Water Resour. Res. 51 (7), 5004–5022.

Del Giudice, D., Reichert, P., Albert, C., Bareš, V., Rieckermann, J., 2015b. Model bias and complexity - understanding the effects of structural deficits and input errors on runoff predictions. Environ. Model. Software.

Dietzel, A., Reichert, P., 2014. Bayesian inference of a lake water quality model by emulating its posterior density. Water Resour. Res. 50 (10), 7626–7647.

Dotto, C.B.S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D.T., Fletcher, T.D., 2011. Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data. Environ. Model. Software 26 (10), 1225–1239.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. 50 (3), 2350–2375.

Forrest, D.R., Hetland, R.D., DiMarco, S.F., 2011. Multivariable statistical regression models of the areal extent of hypoxia over the Texas–louisiana continental shelf. Environ. Res. Lett. 6 (4), 045002.

Freni, G., Mannina, G., Viviani, G., 2009. Urban runoff modelling uncertainty: comparison among Bayesian and pseudo-Bayesian methods. Environ. Model. Software 24 (9), 1100–1111.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. http://www.sciencedirect.com/science/article/pii/S0022169409004843.

Heidelberg University NCWQR, 2015. Accessed Feb 8, 2017. URL https://www.heidelberg.edu/academics/research-and-centers/national-center-for-water-quality-research.

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resour. Res. 49 (8), 4866–4884.

Jiang, S., Jomaa, S., Büttner, O., Meon, G., Rode, M., 2015. Multi-site identification of a distributed hydrological nitrogen model using bayesian uncertainty analysis. J. Hydrol. 529, 940–950.

Kavetski, D., Fenicia, F., Clark, M.P., 2011. Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: insights from an experimental catchment. Water Resour. Res. 47 (5).

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. J. Roy. Stat. Soc. B 63 (3), 425–464.

Menne, M., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R., Gleason, B., et al., 2012. Global Historical Climatology Network-daily (GHCN-daily), Version 3. NOAA National Climatic Data Center.

Muenich, R.L., Kalcic, M.M., Winsten, J., Fisher, K., Day, M., O'Neil, G., Wang, Y.-C., Scavia, D., 2017. Pay-for-performance conservation using swat highlights need for field-level agricultural conservation. Transact. ASABE 60 (6), 1925–1937.

Razavi, S., Tolson, B.A., 2013. An efficient framework for hydrologic model calibration on long data periods. Water Resour. Res. 49 (12), 8418–8431.

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. Water Resour. Res. 45 (10), W10402.

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resour. Res. 48 (9), W09543.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: characterizing rainfall errors using conditional simulation. Water Resour. Res. 47 (11), W11516.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and

predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46 (10), W10531.
Sikorska, A., Del Giudice, D., Banasik, K., Rieckermann, J., 2015. The value of streamflow data in improving TSS predictions—bayesian multi-objective calibration. J. Hydrol. 530, 241–254.
Sorooshian, S., Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: correlated and heteroscedastic error cases. Water Resour. Res. 16 (2), 430–442.
Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour. Res. 39 (8).
Wang, Q., Shrestha, D., Robertson, D., Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resour. Res. 48 (5), W05514.
Westra, S., Thyer, M., Leonard, M., Kavetski, D., Lambert, M., 2014. A strategy for diagnosing and interpreting hydrological model nonstationarity. Water Resour. Res. 50 (6), 5090–5113.
Wooldridge, J.M., 2015. Introductory Econometrics: a Modern Approach. Nelson Education.
Yang, J., Reichert, P., Abbaspour, K., 2007a. Bayesian uncertainty analysis in distributed hydrologic modeling: a case study in the Thur river basin (Switzerland). Water Resour. Res. 43 (10), W10401.
Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007b. Hydrological modelling of the Chaohe basin in China: statistical model formulation and Bayesian inference. J. Hydrol. 340 (34), 167–182.
Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. J. Hydrol. 181 (1–4), 23–48.