Journal of Hydrology 612 (2022) 127968

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Confidence intervals of the Kling-Gupta efficiency

Jasper A. Vrugt *, Debora Y. de Oliveira

Department of Civil and Environmental Engineering, University of California, Irvine, California, USA

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief.

Keywords:
Kling-Gupta efficiency
Nash–Sutcliffe efficiency
Confidence intervals
Bootstrap method
Bayesian analysis
Marginal distribution

ABSTRACT

The Kling-Gupta efficiency, hereafter referred to as KG efficiency rather than its common abbreviation KGE, proposed by Gupta et al. (2009) has become a widely used metric for evaluating the goodness-of-fit of n-vectors of observations, $\tilde{\mathbf{y}} = [\tilde{y}_1\ \tilde{y}_2 \dots \tilde{y}_n]^\top$, and corresponding model simulations, $\mathbf{y}(\boldsymbol{\theta}) = [y_1(\boldsymbol{\theta})\ y_2(\boldsymbol{\theta}) \dots y_n(\boldsymbol{\theta})]^\top$. This metric rectifies some of the shortcomings of the coefficient of determination, R^2, also known to hydrologists as the efficiency of Nash and Sutcliffe (1970), by using a Euclidean-distance based weighting of the correlation, bias and temporal variability of the observed, $\tilde{\mathbf{y}}$, and simulated, $\mathbf{y}(\boldsymbol{\theta})$, data. But as the KG efficiency is not borne out of assumptions with respect to the statistical distribution of the residuals, $\mathbf{e}(\boldsymbol{\theta}) = \tilde{\mathbf{y}} - \mathbf{y}(\boldsymbol{\theta})$, we cannot formally characterize its uncertainty. The NS efficiency suffers a similar problem, yet, statistical theory postulates that its confidence intervals should follow a beta distribution in certain special cases. Without a formal description of the confidence intervals of the KG efficiency, we cannot (amongst others) quantify parameter uncertainty, compute confidence and prediction limits on simulated model responses, inform decision makers about critical modeling uncertainties, evaluate model adequacy and assess the information content of calibration data. More fundamentally, without confidence intervals we cannot establish whether the KG efficiency is a consistent, efficient and unbiased estimator. In this paper we present an empirical description of the confidence intervals of the KG efficiency. We relate the unknown probability distribution of the KG efficiency to the measurement errors of the training data record, $\tilde{\mathbf{y}}$, and use the bootstrap method to carry out statistical inference. We illustrate our method by application to a simple linear regression function for which the least squares parameter confidence regions are exactly known and two hydrologic models of contrasting complexity. The empirical parameter confidence regions and/or intervals of the KG efficiency are compared to those derived from generalized least squares, objective function contouring and Bayesian analysis using Markov chain Monte Carlo simulation. The marginal parameter distributions of the KG efficiency are generally well described by a normal distribution. Results further confirm that the distribution of the KG efficiency is a complex function of data length and the magnitude, distribution and structure of the measurement errors. This prohibits an analytic description of the empirical confidence regions and/or intervals of the KG efficiency and reiterates the need for the bootstrap method.

1. Introduction and scope

Consider a dynamic system model, $\mathscr{M}(\boldsymbol{\theta}, \mathbf{X}) : \mathbb{R}^d \rightarrow \mathbb{R}^n$, which simulates a n-record, $\mathbf{y} = [y_1\ y_2 \dots y_n]^\top$, of a single output variable for a d-vector of parameter values, $\boldsymbol{\theta} = [\theta_1\ \theta_2 \dots \theta_d]^\top$, with $\boldsymbol{\theta} \in \mathbb{R}^p$, and array, $\mathbf{X}$, of constants and input variables required under the supposition or hypothesis that they govern, by causality using physical laws of nature (e.g. mathematical function(s)), the simulated output. The array, $\mathbf{X}$, may characterize the system's initial state and/or invariant (distributed) properties and document the evolution of its spatiotemporal control inputs (forcing/explanatory variables), but is of no particular interest here. Therefore, we suppress use of this symbol and write instead, $\mathbf{y} = \mathscr{M}(\boldsymbol{\theta})$, for the vector-valued form of the model with respect to $\boldsymbol{\theta}$.

A key task is now to determine suitable values of the parameters, $\boldsymbol{\theta}$, so that the model output, $\mathbf{y}$, approximates as closely and consistently as possible the observed system behavior, $\tilde{\mathbf{y}} = [\tilde{y}_1\ \tilde{y}_2 \dots \tilde{y}_n]^\top$. We may now write, $\tilde{\mathbf{y}} = \mathscr{M}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2 \dots \epsilon_n]^\top$ signifies a $n \times 1$ vector of measurement errors. The common paradigm in the statistical literature is to hypothesize a measurement error distribution, $P_n(\boldsymbol{\epsilon})$, of the data, $\tilde{\mathbf{y}}$, and exploit this assumption in the construction of an objective function, $F(\boldsymbol{\theta})$. For example, if the measurement errors satisfy the so-called Gauss–Markov assumptions and (i) have a zero mean, $\mathbb{E}(\epsilon_i) = 0$, (ii)

* Corresponding author.
E-mail address: jasper@uci.edu (J.A. Vrugt).

https://doi.org/10.1016/j.jhydrol.2022.127968
Received 6 July 2021; Received in revised form 17 May 2022; Accepted 20 May 2022
Available online 28 May 2022

constant variance, $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$, for all $i \in \mathbb{N}_+$, and (iii) are uncorrelated, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, for all $i \neq j$, then minimization of the well-known sum of squared residuals

$$F_{\mathrm{SSR}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\widetilde{y}_i - y_t(\boldsymbol{\theta}))^2, \tag{1}$$

will lead to minimum variance estimates of the parameters, $\boldsymbol{\theta}$. The SSR is strictly positive, meaning that $F_{\mathrm{SSR}}(\boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \in \mathbb{R}^p$, possibly with exception of one or more points at which $F_{\mathrm{SSR}}(\boldsymbol{\theta}) = 0$. As complex systems do not admit a perfect characterization, the residuals between model and data, $e_i(\boldsymbol{\theta}) = \widetilde{y}_i - y_i(\boldsymbol{\theta})$ for all $i = (1,2,\ldots,n)$, will, on average, be substantially larger than the data measurement errors, ϵ_i (Gupta et al., 1998; Beven and Binley, 1992; Beven, 2006; Kavetski et al., 2006; Vrugt and Beven, 2018). Nonetheless, the residuals are expected to absorb the consequences of model misspecification, an inadequate characterization of the system properties and errors in the forcing/explanatory variables and initial states and behave statistically in a similar way as the measurement errors.

In this paper we distinguish between formal and informal measures of the goodness-of-fit. To clarify this terminology, formal measures of the goodness-of-fit such as the $F_{\mathrm{SSR}}(\boldsymbol{\theta})$ are the result of the rigorous application of statistical theory and demand explicit (and testable!) assumptions about the probabilistic properties of the residuals. These hypotheses can be verified aposteriori using regression diagnostics. This involves the use of statistical tests for (i) variance homogeneity (Goldfeld and Quandt, 1965; Breusch and Pagan, 1979; White, 1980), (ii) serial correlation (Durbin and Watson, 1950; Durbin and Watson, 1951; Breusch, 1978) and (iii) normality (Anderson and Darling, 1952; Shapiro and Wilk, 1965) of the residuals. These diagnostic checks provide guidance on further stages of model development, hence, constitute an important advantage of the use of formal goodness-of-fit measures. Informal measures of the goodness-of-fit on the contrary do not make assumptions about the expected structure and/or distribution of the residuals. Examples include the coefficient of determination, R^2, better known to hydrologists as the Nash–Sutcliffe (NS) efficiency (Nash and Sutcliffe, 1970) and the Kling-Gupta (KG) efficiency (Gupta et al., 2009). The use of these metrics has profound consequences, the most important of which for this paper is that the lack of clarity in the assumptions about the expected distribution of the residuals prohibits an objective characterization of the confidence and prediction intervals of the parameters and simulated output.

In the past decades, a large number of goodness-of-fit measures have been used to fit hydrologic models to data. This includes the use of (a) formal objective and/or likelihood functions that result from the application of first principles with respect to the statistical properties of the residuals such as the SSR in Eq. (1) and weighted formulations thereof with/without treatment of serial correlation and/or heteroscedasticity within the context of weighted and/or generalized least squares (Tasker, 1980; Stedinger and Tasker, 1985; Kavetski et al., 2006; Kavetski et al., 2006), (b) equivalent likelihood functions (Sorooshian and Dracup, 1980; Kuczera, 1983; Bates and Campbell, 2001), possibly augmented with skew and/or kurtosis (Schoups and Vrugt, 2010; Scharnagl et al., 2015; Ammann et al., 2019) within the context of maximum likelihood estimation and Bayesian inference, (c) informal metrics of fit such as the NS (Nash and Sutcliffe, 1970), KG (Gupta et al., 2009; Knoben et al., 2019) and diagnostic (Schwemmle et al., 2021) efficiencies and possible nonparametric variants (Pool et al., 2018) and other improvements (Lamontagne et al., 2020) within the context of model calibration and/or evaluation, (d) pseudo-likelihood functions within the context of the GLUE methodology (Beven and Binley, 1992; Freer et al., 1996; Beven and Freer, 2001), (e) informal statistical measures of the quality-of-fit such as the coefficient of determination and percentage bias within the context of multiple criteria methods (Gupta et al., 1998; Boyle et al., 2000), (f) hydrologic signatures within the context of model diagnostics (Gupta et al., 2008; Yilmaz et al., 2008; Westerberg et al., 2011), (g) summary metrics within the context of approximate Bayesian computation (Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2013) and (h) tolerable ranges within the context of limits of acceptability (Beven, 2006; Vrugt and Beven, 2018), regional sensitivity analysis (Spear and Hornberger, 1980; Spear et al., 2020) and the parameter identification method based on the localization of information (Vrugt et al., 2002). This arrangement in groups should not imply that certain goodness-of-fit metrics are only used within a particular context. For example, the KG efficiency is not only used as objective function for model calibration but also serves its purpose in model diagnostics (e.g. see Rakovec et al. (2019)).

Since the publication by Gupta et al. (2009), the Kling-Gupta (KG) efficiency has become a widely used metric for evaluating the goodness-of-fit of n-vectors of model simulations, $\mathbf{y}(\boldsymbol{\theta}) = [y_1(\boldsymbol{\theta})\, y_2(\boldsymbol{\theta}) \ldots y_n(\boldsymbol{\theta})]^\top$, and corresponding observations, $\widetilde{\mathbf{y}} = [\widetilde{y}_1\, \widetilde{y}_2 \ldots \widetilde{y}_n]^\top$. This metric rectifies some of the shortcomings of the popular Nash–Sutcliffe (NS) efficiency by using a different, Euclidean-distance based, weighting of the correlation, bias and temporal variability of the observed and simulated data. But this adjustment to the weights of the hydrograph descriptors does not solve the fundamental problem of how to characterize the confidence and prediction limits of the KG efficiency. We face a similar problem with the NS efficiency, yet, for certain special cases we can construct its confidence intervals using the beta distribution (Draper and Smith, 1998) albeit this is rarely done in the literature. Without a description of the confidence intervals of the KG efficiency, we cannot (amongst others) quantify parameter uncertainty, derive confidence and prediction limits on simulated model responses, assess regional relationships between model parameters and catchment characteristics (Kuczera and Parent, 1998), inform decision makers about critical modeling uncertainties, evaluate model adequacy and assess the information content of calibration data (Vrugt et al., 2006). More fundamentally, without confidence intervals we cannot establish whether the KG efficiency is a consistent, efficient and unbiased estimator.

This paper is concerned with the empirical[1] description of the confidence intervals of the KG efficiency. As this estimator lacks a fundamental basis in statistical regression theory, we relate the (distribution of the) KG efficiency to the measurement errors of the training data record, $\widetilde{\mathbf{y}}$, and resort to the bootstrap method of Efron (1979) to characterize its confidence and prediction intervals. Section 2 presents a small sample correction in the decomposition of the NS efficiency of Gupta et al. (2009) as prerequisite to the exact definition of the KG efficiency. This is followed by Section 3 which describes our methodology for estimating the uncertainty of the KG efficiency. We frame our methodology within the context of generalized least squares and alternate between theory and examples involving simple linear regression and nonlinear regression of the rainfall-discharge transformation. We compare the confidence limits of the KG efficiency against those derived from least squares estimation. Finally, Section 4 concludes our paper with a summary of the main findings.

2. The mean squared residual

In this section we briefly review the origins of the KG efficiency which has led to its current use. To this end, we first present the now widely known decomposition of the widely used mean squared residual but with proper treatment of the population and sample variances of the measured, $\widetilde{\mathbf{y}}$, and simulated, $\mathbf{y}(\boldsymbol{\theta})$, data.

[1] We use the wording *empirical confidence intervals* to emphasize the heuristic nature of the confidence intervals of informal goodness-of-fit metrics such as the KG efficiency

2.1. Decomposition

The SSR is intimately related to the mean squared residual[2] or MSR

$$F_{\mathrm{MSR}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{y}_i - y_i(\boldsymbol{\theta})\right)^2, \tag{2}$$

which, as its name suggests, is the average squared difference between the measured and simulated values and, thus, a multiple of $1/n$ of the SSR. Note that the MSR has units equal to the squared dimension of the measurements used. The MSR, in turn, may be normalized by the variance of the training data, $\sigma^2_{\widetilde{y}}$, to yield the well-known coefficient of determination, $R^2 \in (-\infty, 1]$, as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\widetilde{y}_i - y_i(\boldsymbol{\theta})\right)^2}{\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)^2} = 1 - \frac{n\mathrm{MSR}}{n\sigma^2_{\widetilde{y}}} = 1 - \frac{\mathrm{MSR}}{\sigma^2_{\widetilde{y}}}, \tag{3}$$

where $m_{\widetilde{y}}$ signifies the mean of the observed data

$$m_{\widetilde{y}} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{y}_i, \tag{4}$$

and $s^2_{\widetilde{y}}$ is an estimate of the unknown population variance, $\sigma^2_{\widetilde{y}}$

$$s^2_{\widetilde{y}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)^2. \tag{5}$$

As a reminder, it is common practice to use the Greek letters, μ and σ^2, for the theoretical (or population) mean and variance and to use lower case letters, m and s^2, for their sample estimates derived from the data (e.g. see Lamontagne et al. (2020)). This difference in notation is important and consequential for small sample sizes.

For the sample variance, $s^2_{\widetilde{y}}$, to be an unbiased estimator of the true (population) variance, $\sigma^2_{\widetilde{y}}$, of the n-record of measured data, $\widetilde{\mathbf{y}}$, we must divide by $n-1$ rather than n in the denominator of Eq. (5). This is also known as Bessel's correction and a consequence of the use of the sample mean, $m_{\widetilde{y}}$, rather than the (unknown) population mean, $\mu_{\widetilde{y}}$. According to Eq. (5), the sum term, $\sum_{i=1}^{n}(\widetilde{y}_i - m_{\widetilde{y}})^2$, must equal $(n-1)s^2_{\widetilde{y}}$, hence, the coefficient of determination in Eq. (3) may be written as

$$R^2 = 1 - \frac{n\mathrm{MSR}}{(n-1)s^2_{\widetilde{y}}}, \tag{6}$$

The R^2 measures the proportion of the variance of the measured data that is explained by the model, $\mathscr{M}(\boldsymbol{\theta})$. As the denominator of Eq. (3) does not depend on the model output, $\mathbf{y}$, the parameters, $\boldsymbol{\theta}$, that maximize the coefficient of determination, R^2, will minimize the SSR. Nonetheless, the R^2 is classified as an informal metric of quality-of-fit as its definition does not follow from assumptions regarding the nature of the residuals, or more precisely, the measurement errors. The R^2 is also known as the Nash–Sutcliffe efficiency among hydrologists and used widely as measure of model performance of watershed models.

Gupta et al. (2009) has shown that we can decompose the MSR into three different terms

$$\mathrm{MSR} = 2\sigma_y\sigma_{\widetilde{y}}(1-r) + (\sigma_{\widetilde{y}} - \sigma_y)^2 + (m_{\widetilde{y}} - m_y)^2, \tag{7}$$

where $r \in [-1, 1]$ measures the strength of linear association between the measured and simulated data

$$r = \frac{C_{\widetilde{y}y}}{\sqrt{s^2_y}\sqrt{s^2_{\widetilde{y}}}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)\left(y_i - m_y\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - m_y\right)^2}}, \tag{8}$$

which is also known as Pearson's sample correlation coefficient.

2.2. Small sample correction of the NS efficiency

Eq. (7) will return biased values of the MSR if users insert sample variances, $s^2_{\widetilde{y}}$ and s^2_y, of the measured and simulated data, respectively (Lamontagne et al., 2020). We consider this commonplace situation in Appendix A and should use instead the following expression

$$\mathrm{MSR} = \left(\frac{2n-2}{n}\right)s_{\widetilde{y}}s_y(1-r) + \left(\frac{n-1}{n}\right)(s_{\widetilde{y}} - s_y)^2 + (m_{\widetilde{y}} - m_y)^2. \tag{9}$$

The two formulations, Eqs. (7) and (9), are almost similar except for the use of the sample standard deviations, $s_{\widetilde{y}}$ and s_y, of the measured and simulated data and the multiplicative constants, $(2n-2)/n$ and $(n-1)/n$, in front of the first and second term, respectively. These two multipliers are the result of Bessel's correction and yield unbiased estimates of the MSR for small n, say, $n < 20$. The two multiplicative constants approach unity for large n, and, thus, can be removed without harm for multi-year records of daily discharge measured at the catchment outlet.

We can now combine Eqs. (3) and (9) to yield the following expression for the R^2 and, thus, NSE (see Appendix B)

$$R^2 = 2\left(\frac{s_y}{s_{\widetilde{y}}}\right)r_{\widetilde{y}y} - \left(\frac{s_y}{s_{\widetilde{y}}}\right)^2 - \left(\frac{n}{n-1}\right)\left(\frac{m_{\widetilde{y}} - m_y}{s_{\widetilde{y}}}\right)^2 = \mathrm{NSE} = 2\alpha r - \alpha^2 - c\beta_{\mathrm{n}}^2. \tag{10}$$

A similar decomposition of the NSE into three components of correlation, conditional/unconditional bias and/or relative variability was presented by Murphy (1988) and Gupta et al. (2009) but assume knowledge of the (unknown) population means and variances of the measured and simulated data, $\widetilde{\mathbf{y}}$ and $\mathbf{y}(\theta)$, respectively. As a result, our decomposition in Eq. (10) adds a unitless multiplier, $c = n/(n-1)$, to the third term of Eq. (4) of Gupta et al. (2009). The dimensionless scalars, $\alpha > 0$ and $\beta_{\mathrm{n}} \in \mathbb{R}$, measure the relative variability in the simulated and observed values and the normalized bias, respectively

$$\alpha = \frac{s_y}{s_{\widetilde{y}}} \quad \text{and} \quad \beta_{\mathrm{n}} = \frac{m_{\widetilde{y}} - m_y}{s_{\widetilde{y}}}, \tag{11}$$

and $r \in [-1, 1]$ is the well-known sample correlation coefficient of Pearson in Eq. (8). If, instead, we work with the statistical definition, δ (-), of bias

$$\delta = \frac{m_{\widetilde{y}} - m_y}{m_{\widetilde{y}}} \tag{12}$$

then we yield the sample equivalent of the theoretical definition of the efficiency, E, of Lamontagne et al. (2020) as follows

$$R^2 = 2\alpha r - \alpha^2 - cC_{\widetilde{y}}^2\delta^2 = \mathrm{NSE} \tag{13}$$

where $C_{\widetilde{y}} = s_{\widetilde{y}}/m_{\widetilde{y}}$ is the well known coefficient of variation of the measured data and, again, $c = n/(n-1)$, is a testament to Bessel's correction. The use of the standardized bias in Eq. (13) simplifies

[2] The definition mean squared error (MSE) is widespread in the literature, but inconsistent when its computation involves unobservable errors. The word error implies a difference between an observed value and its true value. As our measurements of system behavior are imperfect, the residuals, $\mathbf{e}(\boldsymbol{\theta}) = \widetilde{\mathbf{y}} - \mathbf{y}(\boldsymbol{\theta})$, are estimates of the errors under the assumed model, $\mathbf{y} = \mathscr{M}(\boldsymbol{\theta})$. Hence, we should use the word residual instead.

comparison across models and/or watersheds (Lamontagne et al., 2020).

2.3. The Kling-Gupta efficiency

Drawing inspiration from the decomposition of the NSE, Gupta et al. (2009) proposed a new criterion, the so-called Kling-Gupta (KG) efficiency, to read

$$\mathrm{KG} = 1 - \mathrm{ED}, \tag{14}$$

where ED equals the Euclidean distance of (α, β, r) to the so-called ideal point, $(1, 1, 1)$, and is computed using

$$\mathrm{ED} = \sqrt{(\alpha-1)^2 + (\beta-1)^2 + (r-1)^2}, \tag{15}$$

where $\beta = m_y/m_{\widetilde{y}}$ is a unitless measure of the bias ($\beta \in \mathbb{R}$). Like the R^2, and, thus NSE, the KG efficiency can take on values between minus infinity and one, where unity implies a perfect fit to the measured data. Gupta et al. (2009) has demonstrated the advantages of the KG efficiency over the widely used NSE. This has stimulated widespread use of the KG efficiency in hydrology for evaluating the goodness-of-fit between model simulations, $\mathbf{y}$, and corresponding observations, $\widetilde{\mathbf{y}}$. One potential concern with the application of the KG efficiency is that the three components, α, β and r, of Eq. (15) are ratios of product moments that are known to exhibit enormous bias for skewed data such as daily streamflow records (Lamontagne et al., 2020; Vogel and Fennessey, 1993; Barber et al., 2019). This problem persists even for long streamflow time series and should be avoided. Lamontagne et al. (2020) presents improved estimators of the NSE and KG efficiency in controlled Monte Carlo experiments.

This paper is not concerned with a formal mathematical analysis of the empirical sampling properties of the KG efficiency as in Lamontagne et al. (2020) but rather focuses attention on the confidence intervals of this estimator. As the KG efficiency is an informal measure of the goodness-of-fit, its optimal parameter estimates, $\boldsymbol{\theta}^*$, that maximize Eq. (14) do not enjoy a formal description of their confidence intervals. This impairs our ability to quantify model parameter and predictive uncertainty of the KG efficiency. In fact, most hydrologic signatures that are used in watershed model diagnostics suffer a similar limitation. Certainly, we would favor a statistical description of the empirical distribution of the KG efficiency. A quantitative description of the empirical confidence intervals of the KG efficiency serves many practical tasks and purposes of which uncertainty quantification of model parameter and simulated output is most important from the perspective of this paper. More fundamentally, this knowledge of the confidence intervals of the KG efficiency is a necessary requirement for a formal analysis of the asymptotic behavior, consistency, efficiency and unbiasedness of this estimator.

3. Empirical description of uncertainty of the Kling-Gupta efficiency

We present our empirical description of the uncertainty of the KG efficiency. We frame our methodology within the context of generalized least squares and alternate between theory and ensuing case studies using parameter estimation problems in linear and nonlinear regression. This order of presentation helps to convey our arguments and methodology and should help readers understand how current developments relate to statistical regression theory. Certainly, we should not ignore and/or forget about least squares methods in our efforts to push forward the envelope in hydrologic model calibration and evaluation.

3.1. Linear regression

3.1.1. Theory

To clarify our approach, we write the measurements of the training record as follows

$$\widetilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\epsilon}, \tag{16}$$

where $\mathbf{y} = [y_1 \ \dots \ y_n]^\top$ is the "true" response of the data generating system, $\Im$. We would want our simulated output to mimic as closely and consistently as possible this unobserved, measurement error-free, response. To help uncover and replicate the unobserved true response, $\mathbf{y}$, we must make some assumptions about the nature and distribution of the measurement errors, $\boldsymbol{\epsilon}$. When the ϵ_i's in Eq. (16) are expected to have a zero-mean with constant variance, σ_ϵ^2, then the MSR will provide a meaningful assessment of model performance. This constancy assumption, however, does not do justice to variables such as river discharge whose measurement errors increase with the measured flow level, $\widetilde{\mathbf{y}}$, and, thus, $\epsilon_i \widetilde{\sim} \mathcal{N}(0, \sigma_{\epsilon_i}^2) \ \forall i \in (1, \dots, n)$, where the symbol, $\widetilde{\sim}$, means *distributed according to*.

The magnitude and structure of the measurement errors of the training data, $\widetilde{\mathbf{y}}$, will exert control on the quality of the parameter estimates, hence, this information should be incorporated in the MSR estimator for a meaningful comparison of the measured and modeled system response. This extension of the MSR to correlated and/or heteroscedastic measurement errors is also known as generalized least squares and was first described by Aitken (1936). He showed that if the measurement errors, $\boldsymbol{\epsilon} = [\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_n]^\top$, have (i) a zero mean, $\mathbb{E}(\epsilon_i) = 0 \ \forall i = (1, 2, \dots, n)$ and (ii) covariance matrix, $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) = \Sigma_\epsilon$, then minimization of the generalized least squares (GLS) objective function

$$F_{\mathrm{GLS}}(\boldsymbol{\theta}) = \mathbf{e}(\boldsymbol{\theta})^\top \Sigma_\epsilon^{-1} \mathbf{e}(\boldsymbol{\theta}), \tag{17}$$

produces unbiased and minimum variance estimates of the parameters, $\boldsymbol{\theta}$. The symbol $^\top$ denotes matrix transpose and turns the $n \times 1$ residual vector, $\mathbf{e}(\boldsymbol{\theta})$, into a row vector so as to have matching inner dimensions in Eq. (17). The diagonal entries of the $n \times n$ measurement error covariance matrix, Σ_ϵ, specify the variances of the measurement errors of the $\widetilde{y}_i$'s, while the off-diagonal entries list the pairwise covariances of ϵ_i and ϵ_j for all $i, j \in (1, 2, \dots, n)$ and $i \neq j$. To better understand the inner workings of the vector-matrix–vector product of Eq. (17) we define the $n \times n$ weight matrix, $\mathbf{W}$, to be the square root of Σ_ϵ^{-1}, so that $\mathbf{W} = \Sigma_\epsilon^{-\frac{1}{2}}$, and, thus, $\mathbf{W}^\top \mathbf{W} = \Sigma_\epsilon^{-1}$. Then, Eq. (17) can be written as a vector inner product, $F_{\mathrm{GLS}}(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}(\boldsymbol{\theta})^\top \boldsymbol{\varepsilon}(\boldsymbol{\theta})$, of the homogenized and/or decorrelated residuals, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = \mathbf{W}\mathbf{e}(\boldsymbol{\theta})$. The entries of the $n \times 1$ vector $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$ are also referred to as partial residuals or innovations.

When the measurement errors are known to be uncorrelated, the off-diagonal entries of Σ_ϵ and, thus, $\mathbf{W}$, are zero and Eq. (17) specializes to a weighted sum of squared residuals (WSSR) objective function

$$F_{\mathrm{WSSR}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(\widetilde{y}_i - y_i(\boldsymbol{\theta}))^2}{\sigma_{\epsilon_i}^2} = \sum_{i=1}^{n} w_i^2 e_i(\boldsymbol{\theta})^2 = \sum_{i=1}^{n} \varepsilon_i(\boldsymbol{\theta})^2, \tag{18}$$

where the weights, $w_i = 1/\sigma_{\epsilon_i} \ \forall i \in (1, \dots, n)$, on the main diagonal of $\mathbf{W}$ are equal to the reciprocal of the measurement error standard deviations and $\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = [\varepsilon_1(\boldsymbol{\theta}) \ \varepsilon_2(\boldsymbol{\theta}) \ \dots \ \varepsilon_n(\boldsymbol{\theta})]^\top$ is the $n \times 1$ vector of homogenized (partial) residuals. In the case of homoscedastic errors, the weights, w_i, are all equal, hence, we can write, $\Sigma_\epsilon = \sigma_\epsilon^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of size n and the WSSR reduces to a multiple of $1/\sigma_\epsilon^2$ of the SSR.

The use of the $n \times n$ covariance matrix of the measurement errors, Σ_ϵ, places the GLS and, thus, WSSR, objective functions on a firm statistical footing. Suppose that we would like to fit the linear regression function, $\widetilde{\mathbf{y}} = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, with $n \times p$ design matrix, $\mathbf{D}$, and $p \times 1$ parameter vector, $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \dots \ \theta_p]^\top$, to the n entries, $\widetilde{y}_1, \widetilde{y}_2, \dots, \widetilde{y}_n$, of the training data record, $\widetilde{\mathbf{y}}$, with $n \times n$ measurement error covariance matrix, Σ_ϵ. The GLS parameter values, $\widehat{\boldsymbol{\theta}}$, can be obtained by setting the derivative of Eq. (17) with respect to $\boldsymbol{\theta}$ to zero to yield

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \vdots \\ \widehat{\theta}_p \end{bmatrix} = \left(\mathbf{D}^\top {\Sigma_\epsilon}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^\top {\Sigma_\epsilon}^{-1}\tilde{\mathbf{y}}. \tag{19}$$

where $\mathbf{C}(\widehat{\boldsymbol{\theta}}) = \left(\mathbf{D}^\top {\Sigma_\epsilon}^{-1}\mathbf{D}\right)^{-1}$ equals the $p \times p$ variance–covariance matrix of the GLS parameters. This positive-definite matrix $p \times p$ matrix defines an elliptical distribution, a generalization of the multivariate normal distribution, with density function

$$p(\boldsymbol{\theta}) = k\psi\left((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{C}(\widehat{\boldsymbol{\theta}})^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right) \tag{20}$$

where k is a normalization constant and the scalar function, $\psi(x)$, returns the unnormalized density at x. The 100γ% confidence region of the GLS parameter values, is now made up all parameter vectors, $\boldsymbol{\theta} \in \mathbb{R}^p$, that lie inside the space delineated by

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{C}(\widehat{\boldsymbol{\theta}})^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leqslant f(p_\gamma, p, n-p), \tag{21}$$

at the critical value, $p_\gamma = \gamma$. The above inequality describes an ellipse for $p = 2$, an ellipsoid in three dimensions and a hyperboloid in p-dimensional Cartesian parameter space, $\mathbb{R}^p$, although other shapes can occur. The principal axes (volume, shape) and orientation (direction, angle) of the confidence region of the hyperboloid are determined by the inverse parameter covariance matrix, also referred to as Fisher information matrix, $\mathcal{J}(\boldsymbol{\theta})$, after Sir Ronald Aylmer Fisher (1890–1962). The larger the values of $\mathcal{J}(\boldsymbol{\theta}) = \mathbf{D}^\top {\Sigma_\epsilon}^{-1}\mathbf{D}$, the stronger the curvature of $F_{\mathrm{GLS}}(\boldsymbol{\theta})$, and the smaller the uncertainty of the parameters (and, thus, volume of the confidence region).

To link our assumptions about the probabilistic properties of the measurement errors, ϵ, to the GLS parameters, $\widehat{\boldsymbol{\theta}}$, we write the $n \times n$ covariance matrix of the measurement errors, $\Sigma_\epsilon = \sigma_\epsilon^2\mathbf{V}$, and, thus, $\epsilon \overset{\mathscr{D}}{\sim} \mathcal{N}_n\left(\mathbf{0}, {\sigma_\epsilon}^2\mathbf{V}\right)$. In the case of homoscedastic and uncorrelated measurement errors, $\mathbf{V} = \mathbf{I}_n$, otherwise, the nonsingular $n \times n$ matrix $\mathbf{V}$ may have unequal diagonal elements (heteroscedasticity) and/or off-diagonal entries that are non-zero. The $p \times p$ parameter covariance matrix, $\mathbf{C}(\boldsymbol{\theta})$, can now be written as follows

$$\mathbf{C}(\boldsymbol{\theta}) = s_\epsilon^2\left(\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}\right)^{-1}, \tag{22}$$

with corresponding geometric description of the 100γ% confidence region

$$\frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{{s_\epsilon}^2} \leqslant f(p_\gamma, p, n-p), \tag{23}$$

where s_ϵ^2 is the sample variance of the weighted (homogenized and/or decorrelated) residuals

$$s_\epsilon^2 = \frac{\mathbf{e}(\widehat{\boldsymbol{\theta}})^\top\mathbf{V}^{-1}\mathbf{e}(\widehat{\boldsymbol{\theta}})}{n-p}. \tag{24}$$

According to the definition of the chi-square distribution, the sum of squares of the weighted residuals in the numerator of Eq. (24) will follow a multiple, σ_ϵ^2, of the χ^2-distribution with $n-p$ degrees of freedom. This implies that $s_\epsilon^2 \overset{\sim}{\sim} {\sigma_\epsilon}^2\chi^2_{n-p}/(n-p)$, with loss of one degree of freedom for each parameter of the regression function. Analogously, the sum of squares of p weighted parameter deviations from the center, $\widehat{\boldsymbol{\theta}}$, of the ellipsoidal region defined in the numerator of Eq. (23) will follow a multiple, σ_ϵ^2, of the chi-square distribution with p degrees of freedom, thus, $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{\mathscr{D}}{\sim} {\sigma_\epsilon}^2\chi_p^2$. If we put everything together, we yield

$$\frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{s_\epsilon^2} \overset{\mathscr{D}}{\sim} \frac{\sigma_\epsilon^2\chi_p^2}{\sigma_\epsilon^2\chi^2_{n-p}/(n-p)} = \frac{p\chi_p^2/p}{\chi^2_{n-p}/(n-p)}. \tag{25}$$

The ratio of two chi-squared variates, u_1 and u_2, with ν_1 and ν_2 degrees of freedom

$$X = \frac{u_1/\nu_1}{u_2/\nu_2}, \tag{26}$$

produces a variate, X, which follows a $\mathscr{F}$-distribution, $\mathscr{F}(\nu_1, \nu_2)$ with ν_1 and ν_2 degrees of freedom. Thus, the joint 100γ% confidence region of the GLS parameters, $\widehat{\boldsymbol{\theta}}$, now satisfies

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leqslant ps_\epsilon^2F^{-1}_{\mathscr{F}}\left(p_\gamma\middle|p, n-p\right), \tag{27}$$

where $F^{-1}_{\mathscr{F}}(p_\gamma\middle|p,\nu)$ signifies the inverse of the Fisher-Snedecor cumulative distribution function (cdf) with p and $\nu = n-p$ degrees of freedom at the critical value, $p_\gamma = \gamma$. The inverse cdf is also called the quantile or percent-point function and returns the value, x, of random variable, X, at which $P(X \leqslant x) = p_\gamma$. In other words, the inverse cdf, $F^{-1}_{\mathscr{F}}(p_\gamma\middle|\cdot)$, of some univariate distribution, $\mathscr{X}(\cdot)$, returns the unique real number, x, so that $F_{\mathscr{X}}(x\middle|\cdot) = p_\gamma$. For long training data records, $\tilde{\mathbf{y}}$, the distribution of the sample variance of the weighted residuals in the denominator of Eq. (25) reduces to σ_ϵ^2 and the 100γ% confidence region becomes

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leqslant ps_\epsilon^2F^{-1}_{\chi^2}\left(p_\gamma\middle|p\right). \tag{28}$$

Multivariate confidence regions are difficult to visualize, and, thus, pseudo-univariate intervals may be determined instead from the diagonal entries of the $p \times p$ parameter covariance matrix, $\mathbf{C}(\widehat{\boldsymbol{\theta}})$, as follows

$$\boldsymbol{\theta}_\gamma = \widehat{\boldsymbol{\theta}} \pm \sqrt{\mathrm{Diag}(\mathbf{C}(\widehat{\boldsymbol{\theta}}))}F^{-1}_{\mathscr{T}}\left(\frac{1}{2}(1-\gamma)\middle|n-p\right), \tag{29}$$

where $F^{-1}_{\mathscr{T}}(p_\gamma\middle|\nu)$ is the inverse of the Student's t-cumulative distribution function at cumulative probability (percentile), $p_\gamma = \frac{1}{2}(1 \pm \gamma)$, and degrees of freedom, $\nu = n-p$. These confidence intervals are projections of the confidence region on individual parameter axes. For $\gamma = 0.95$ we yield a 95% parameter confidence interval and the critical t-value, $F^{-1}_{\mathscr{T}}(p_\gamma\middle|\nu)$, equals $12.71, 2.57$ and 1.96 for $n = 1, n = 5$ and $n \to \infty$, respectively.

Confidence limits of the least squares simulated output, $\widehat{\mathbf{y}} = \mathbf{D}\widehat{\boldsymbol{\theta}} = \left[\widehat{y}_1\ \widehat{y}_2 \ldots \widehat{y}_n\right]^\top$, of the linear regression function, $\tilde{\mathbf{y}} = \mathbf{D}\boldsymbol{\theta} + \epsilon$, can be computed as follows

$$\widehat{y}_{i,\gamma} = \widehat{y}_i \pm F^{-1}_{\mathscr{T}}\left(p_\gamma\middle|\nu\right)s_\epsilon\sqrt{\mathbf{d}_i^\top\left(\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{d}_i}, \tag{30}$$

where $\mathbf{d}_i^\top$ signifies the i^{th} row of the design matrix and $i = (1, 2, \ldots, n)$. The confidence intervals follow a normal distribution with mean equal to the least squares output, $\widehat{y}_i = \mathbf{d}_i\widehat{\boldsymbol{\theta}}$, and variance determined by $\mathbf{d}_i$ and parameter covariance matrix, $\mathbf{C}(\widehat{\boldsymbol{\theta}}) = s_\epsilon^2\left(\mathbf{D}^\top\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$.

3.1.2. Application: linear function

To illustrate the application of Eqs. (27) and (29), please consider Fig. 1a which presents a contour plot of the GLS objective function for a simple regression function, $y_i = f(a, b, t_i) = at_i + b$, using synthetic training data, $\tilde{\mathbf{y}} = \mathbf{y} + \epsilon$, created using $a = 1, b = 2, t = (1, 2, \ldots, 50)$ and measurement errors, $\epsilon \overset{\mathscr{D}}{\sim} \mathcal{N}_n(\mathbf{0}, \Sigma_\epsilon)$, drawn at random from a n-variate normal distribution with zero mean and $n \times n$ covariance matrix, $\Sigma_\epsilon =$

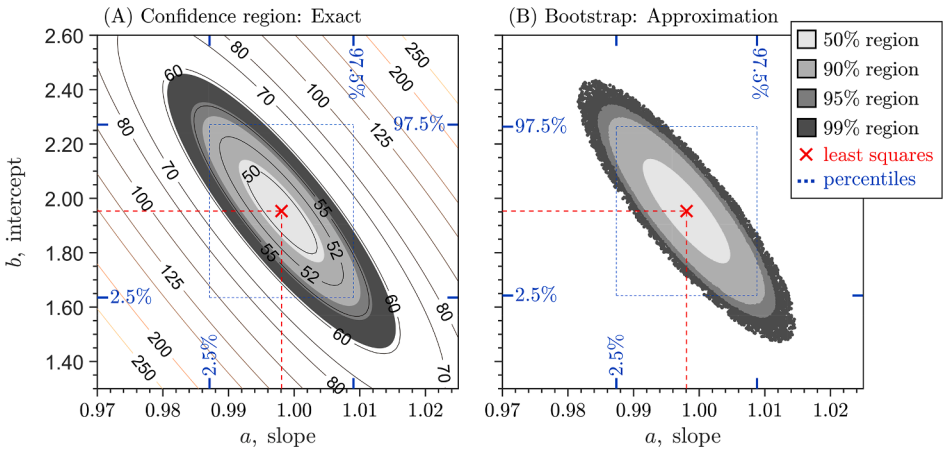

Fig. 1. Illustrative example of parameter confidence regions and confidence intervals: (a) Contour plot of the generalized least squares objective function of Eq. (17) for a linear regression function, $f(\boldsymbol{\theta}, t) = at + b$, using $0.97 \leqslant a \leqslant 1.025$ and $1.30 \leqslant b \leqslant 2.60$. The contour lines are labeled and coded using a copper colormap. The bivariate $100\gamma\%$ confidence regions of the slope, a, and intercept, b, defined by Eq. (27) are displayed with a gray color scheme using $\gamma = 0.50, \gamma = 0.90, \gamma = 0.95$ and $\gamma = 0.99$, respectively. The dashed blue lines display the univariate 95% confidence intervals of the slope and intercept computed from Eq. (29); (b) bivariate scatter plot of the $N = 10,000$ optimized values of a and b derived from the bootstrap method. The color of the dots signifies the confidence level derived from the bootstrap samples. The dashed blue lines portray the univariate 95% intervals of a and b. The red cross highlights the minimum of the GLS objective function.

$\sigma_\epsilon^2 \mathbf{V}$, where $\sigma_\epsilon^2 = \frac{1}{2}$ and $\mathbf{V} = \mathbf{I}_n$. In vector form, the regression function reads, $y_i = \mathbf{d}(\widetilde{t}_i)\boldsymbol{\theta}$, where $\mathbf{d}(\widetilde{t}_i) = [\widetilde{t}_i \; 1]$ signifies the i^{th} row of the $n \times p$ design matrix, $\mathbf{D}$, and $\boldsymbol{\theta} = [a \; b]^\top$. The colored ellipses portray the $100\gamma\%$ confidence regions of a (slope) and b (intercept) using $\gamma = 0.50$ (light gray), $\gamma = 0.90$ (light-medium gray), $\gamma = 0.95$ (medium gray) and $\gamma = 0.99$ (dark gray). The dotted blue lines portray separately the 2.5% and 97.5% percentiles that make up the 95% confidence intervals of the regression model parameters, a and b, derived from Eq. (29). The red plus characterizes the GLS solution of a and b.

The $100\gamma\%$ confidence region of the slope and intercept, a and b, equals a thin ellipse that centers on the least squares solution (red cross). The length and direction of the two principal axes of the ellipse are given by the eigenvalues and eigenvectors, respectively, of the parameter covariance matrix, $\mathbf{C}(\boldsymbol{\theta})$. As a result, the 99% confidence region of the two parameters, a and b, is simply a multiple of their 95% region. The ellipses enclose values of the slope and intercept, (a, b), which the training data, $\widetilde{\mathbf{y}}$, suggests are statistically acceptable at a given confidence level, γ. The univariate 95% confidence intervals of the slope and intercept (blue dotted lines) underestimate their bivariate counterparts, nevertheless, provide a reasonable description of the individual parameter ranges. One should be careful, however, in interpreting these intervals for a and b as a joint confidence region. Indeed, points in the bottom-left or upper-right corners of the rectangular region delineated by the univariate 95% intervals may seem reasonable for (a, b), yet, the joint 95% confidence region of the two parameters as characterized exactly by the ellipse in medium gray, demonstrates that points in the white region are inadequate. In other words, the hypercube defined by expression (29) can be very different from the proper joint confidence region of the parameters. This discrepancy between univariate and multivariate confidence intervals is well known in the statistical literature (Draper and Guttman, 1995), and researchers usually present univariate confidence intervals only. In the general case with $p > 3$ parameters the differences between the marginal and joint confidence regions are typically larger and more difficult to visualize. Therefore, most researchers resort to univariate confidence intervals only. Note that one can change the projection of the confidence regions on the parameter axes and/or adapt the rectangular block to have an equal size (volume) as the ellipsoidal region (Draper and Guttman, 1995). For example, Press et al. (1992) suggests replacing $F_{\mathcal{F}}^{-1}\left(\frac{1}{2}(1-\gamma)\middle|n-p\right)$ in Eq. (29) with $F_{\chi^2}^{-1}(p_\gamma|p)$. This may result in a better agreement of the confidence interval block and its ellipsoidal counterpart(s), yet, for reasons demonstrated herein, this univariate description of parameter uncertainty cannot replace the joint confidence region.

Certainly, the mathematical description of the confidence region in Eq. (27), is only exact for regression functions with valid basis functions. Then, the $n \times p$ design matrix, $\mathbf{D}$, is fixed and the principal axes of the ellipse described by $\mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D}$ are independent of $\boldsymbol{\theta} \in \mathbb{R}^p$. For all other functions, the sensitivity (design) matrix will depend on the parameter values and the confidence region expressed in Eq. (27) is at best only an approximation of the true multivariate parameter uncertainty. This is no grounds for panic, but simply a reason to change the approach to how we construct the confidence regions. Fortunately, the GLS objective function is rooted in statistical theory, and, therefore, we can choose among several different methods to describe exactly the multivariate parameter uncertainty. This includes (among others) (i) contouring of the GLS objective function, (ii) Monte Carlo simulation and (iii) the bootstrap method. These three methods are fundamentally different, but share in common an exhaustive description of the GLS cost function in the neighborhood of $\widehat{\boldsymbol{\theta}}$. As the first of these two methods demand use of formal goodness-of-fit measures and/or likelihood functions which originate from residual assumptions (see Appendix C), in their current form they are not capable of characterizing parameter uncertainty associated with the application of informal quality of fit metrics such as the KG efficiency. Hence, we focus our attention on the bootstrap method as this approach best suits our application as will be demonstrated next.

Fig. 1b illustrates the results of the bootstrap method of Efron (1979) by application to the regression function, $y_i = f(a, b, t_i) = at_i + b$, used herein. The scatter plot visualizes $(\widehat{a}, \widehat{b})$ data pairs derived from the repeated application of Eq. (19) to $N = 10,000$ different realizations, $\widetilde{\mathbf{y}}_{\mathrm{r}} = \widetilde{\mathbf{y}} + \epsilon$, of the training data record, $\widetilde{\mathbf{y}}$, drawn at random from the n-variate normal distribution with mean, $\widetilde{\mathbf{y}}$, and covariance matrix, $\Sigma_\epsilon = s_\epsilon^2 \mathbf{V}$, and, thus, $\widetilde{\mathbf{y}}_{\mathrm{r}} \widetilde{\sim} \mathcal{N}_n(\widetilde{\mathbf{y}}, s_\epsilon^2 \mathbf{V})$. This is equivalent to our formulation of the training data in Eq. (16) with measurement errors, $\epsilon \widetilde{\sim} \mathcal{N}_n(0, s_\epsilon^2 \mathbf{V})$. The variance of the homogenized residuals, s_ϵ^2, is computed from the GLS solution, $\widehat{\boldsymbol{\theta}}$, using Eq. (24). The bootstrap samples are color coded based on their percentiles of the GLS objective function of Eq. (17) computed using the measured training data, $\widetilde{\mathbf{y}}$, not the replicate records. The $100\gamma\%$ confidence region(s) of the slope and intercept derived from the bootstrap method match exactly their analytic counterparts of Eq. (27). This is true for all critical values. The small imperfections in the outside perimeter of the outermost ellipse sampled by the bootstrap method highlights the need for a sufficiently large sample size. This frontier demarcates the edges of the 99% confidence region and is characterized by the most improbable realizations of the training data record. We also witness an excellent agreement between the univariate 95% bootstrap

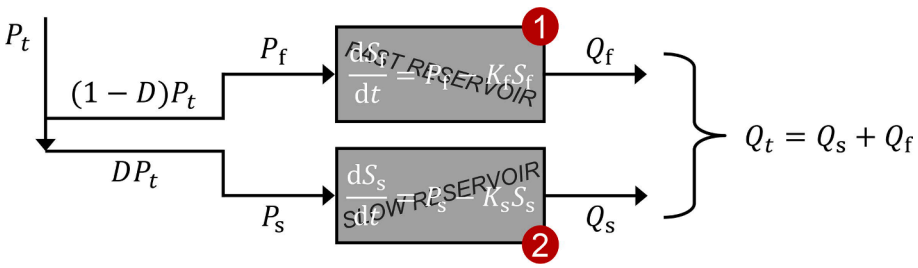

Fig. 2. Schematic illustration of the three parameter hydrologic model. Grey boxes, labeled in red, correspond to fictitious control volumes which control the transformation of rainfall into river discharge. Arrows portray the fluxes into and out of the compartments, including daily precipitation, P_t, the inflows, $(1-D)P_t$ and DP_t, to the fast and slow reservoirs (in mm/d), and the fast, Q_f, and slow, Q_s, reservoir's contribution to the discharge. This simple model admits an analytic solution for the simulated discharge.

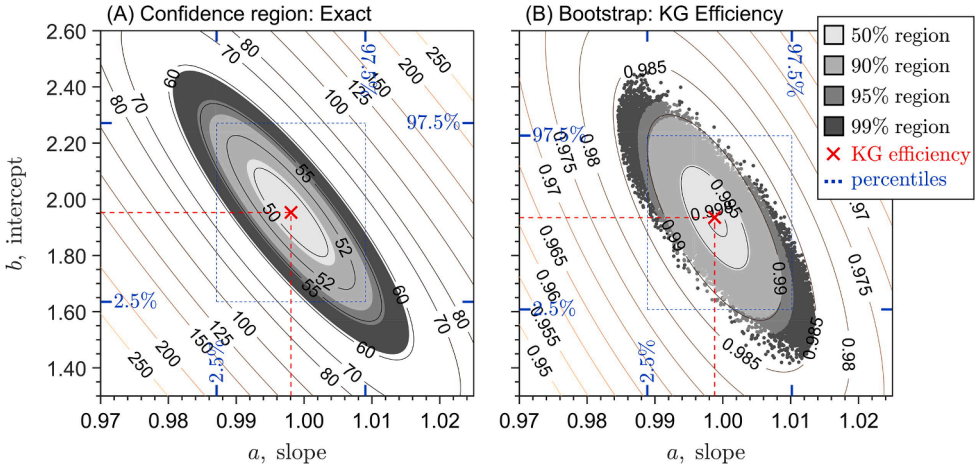

Fig. 3. Bivariate scatter plots of the parameter pairs of the three-parameter hydrologic model, (a) (D,K_f), (b) (K_f,K_s) and (c) (K_s,D), using the bootstrap method (top panel) and DREAM algorithm (bottom panel). The samples are color coded to reveal their respective confidence levels, $\gamma = 0.5$ (light gray), $\gamma = 0.90$ (light-medium gray), $\gamma = 0.95$ (medium gray) and $\gamma = 0.99$ (dark gray). The optimum parameter values are separately indicated in each graph with a red cross.

confidence intervals of a and b and their analytic values derived from Eq. (29).

In summary, if we perturb the measured data, $\tilde{y} = [\tilde{y}_1\ \tilde{y}_2 \dots \tilde{y}_n]^\top$, according to the assumptions of the GLS estimator, then the optimized parameter values for each replicate record, $\tilde{\mathbf{y}}_r$, of the training data, $\tilde{\mathbf{y}}$, define a p-variate distribution with probability density function synonymous to Eq. (20) and 100γ% confidence region(s) described by Eq. (27). This is the underlying idea of approximate Bayesian computation (see e.g. Vrugt and Sadegh (2013)) and allows for an exact description of parameter uncertainty in the absence of convenient closed-form solutions for their confidence regions and/or intervals such as Eqs. (29) and (27) for the GLS, WSSR and/or SSR objective functions. Next, we should confirm that these conclusions also hold for regression functions with invalid basis functions.

3.2. Nonlinear regression

3.2.1. Theory

The regression function, $f(\boldsymbol{\theta},t) = at + b$, of the first case study satisfies the linearity condition

$$f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}, t_i) = f(\boldsymbol{\theta}, t_i) + \mathbf{d}_i^\top \Delta\boldsymbol{\theta}, \tag{31}$$

where $\mathbf{d}_i^\top = [\,t_i\ \ 1\,]$ is the i^{th} row of the $n \times p$ design matrix, $\mathbf{D}$, and $i = (1, 2, \dots, n)$. This linearity condition does not hold for regression functions whose output depends nonlinearly on their parameters. Then, the entries of the design matrix, $\mathbf{D}$, may not only depend on the explanatory variable, t, but also on the values of one or more parameters. This is commonplace in hydrology and has two important implications. The GLS parameter values, $\widehat{\boldsymbol{\theta}}$, cannot be determined from the closed-form solution in Eq. (19) but instead must be estimated using an iterative search and/or optimization method. Furthermore, we should not expect Eqs. (29) and (30) to provide an exact description of the 100γ% confidence intervals of the parameters and simulated output. Next, we illustrate the application of the bootstrap method to nonlinear regression.

3.2.2. Application: a hydrologic toy model

Our second case study considers a simple 3-parameter hydrologic model comprised of two linear reservoirs organized in parallel (see Fig. 2).

This model has two state variables, S_f and S_s, with units of mm, and three parameters, the unitless rainfall distribution coefficient, $D \in (0,1]$, and the recession constants, K_s and K_f, of the slow and fast reservoirs, respectively, with dimensions of reciprocal day. We set, $D = 0.65, K_s = 0.035$ and $K_f = 0.7$ and create a 5-year simulation of daily discharge, $\tilde{y} = [y_1\ y_2 \dots y_n]^\top$ (in mm/day) via an analytic solution using as model input a hypothetical record of daily rainfall data. Each entry of the discharge simulation is subsequently perturbed with a heteroscedastic measurement error, $\epsilon \overset{\mathscr{D}}{\sim} \mathcal{N}_n(\mathbf{0}, \Sigma_\epsilon)$, to yield, $\tilde{y}_t = y_t + \epsilon_t$. The $n \times n$ measurement error covariance matrix, $\Sigma_\epsilon = c\mathbf{V}$, where $c = 0.01$ and the $n \times n$ matrix $\mathbf{V}$ has zeros everywhere except for the main diagonal which lists the squared values of the n simulated discharges. This amounts to a heteroscedastic measurement error with standard deviation equal to 10% of the simulated discharge. We would now like to use the measured data, $\tilde{y} = [\tilde{y}_1\ \tilde{y}_2 \dots \tilde{y}_n]^\top$, to determine the confidence regions of the parameters. As the model does not have valid basis functions, we cannot write the model in matrix form and must use an optimization method to minimize the GLS objective function in Eq. (17). We assume perfect knowledge of the measurement errors, thus, admit Σ_ϵ to our analysis.

Fig. 3 (top panel) presents the results of the bootstrap method using repeated minimization of the GLS objective function in Eq. (17) for replicate samples, $\tilde{\mathbf{y}}_r \tilde{\sim} \mathcal{N}_n(\tilde{\mathbf{y}}, \Sigma_\epsilon)$, of the measured discharge record. The

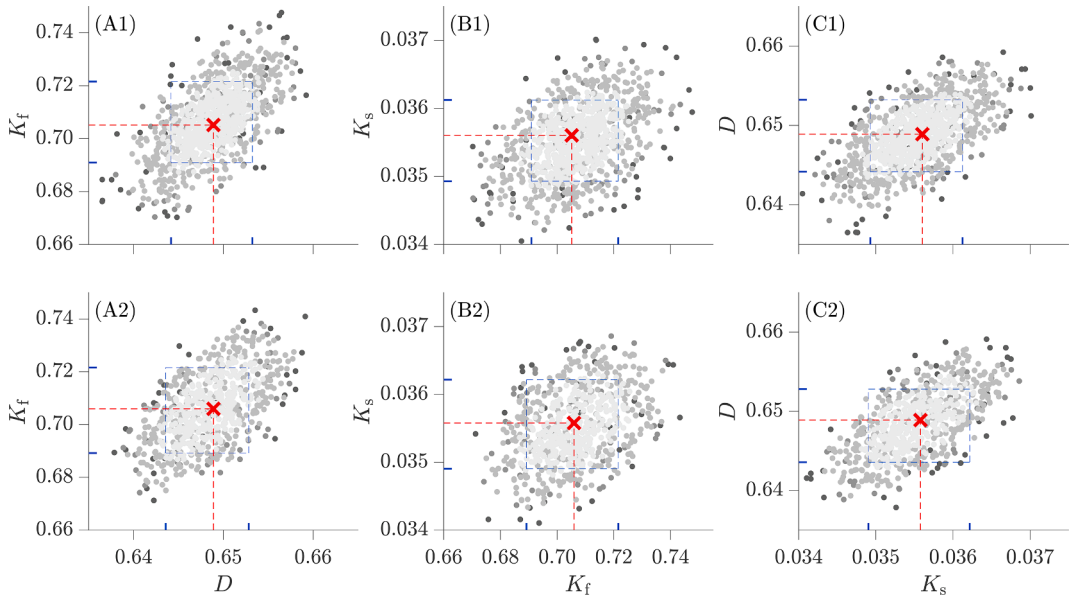

Fig. 4. Visualization of the bivariate confidence region of the slope, a, and intercept, b, of the linear regression function, $f(\boldsymbol{\theta}, t) = at + b$, for (a) the GLS objective function using Eq. (27) and (b) the KG efficiency with the bootstrap method. The confidence regions are coded with a gray color scheme using $\gamma = 0.50, \gamma = 0.90, \gamma = 0.95$ and $\gamma = 0.99$, respectively. The dashed blue lines depict the univariate 95% confidence intervals of the slope and intercept derived from (a) Eq. (29) and (b) the percentiles of the bootstrap samples. The red cross corresponds to (a) the minimum of the GLS objective function and (b) the maximum of the KG efficiency.

gray tints of the bootstrap samples differ based on their confidence levels including, $\gamma = 0.5$ (light gray), $\gamma = 0.90$ (light-medium gray), $\gamma = 0.95$ (medium gray) and $\gamma = 0.99$ (dark gray). The red cross in each graph indicates the GLS optimum. The bottom panel serves as our benchmark and presents scatter plots of the bivariate samples of the posterior distribution derived from the DREAM algorithm (Vrugt et al., 2009). In keeping with the GLS assumptions, we must specify a uniform prior parameter distribution in connection with the likelihood function of Eq. (C4) in Appendix C with $\widehat{\sigma}_\epsilon^2 = 0.01$.

The bivariate confidence regions of the parameter pairs appear symmetric around their optimum value with sampling density that decreases away from the center of the point clouds. The confidence regions are well described by ellipses, and their diagonal orientation suggests the presence of parameter correlation among, D, K_s and K_s. This is all interesting, yet, most important is the nearly perfect match of the confidence regions of the bootstrap method and the DREAM algorithm. This is not a surprise, nevertheless, an important demonstration to those unfamiliar with the bootstrap method and its applicability to uncertainty quantification. It needs no further demonstration that the confidence intervals of both methods are a perfect match.

We are now ready to pair the bootstrap method with the KG efficiency. But before doing so, we first would like to provide some general remarks about the limitations of the bootstrap method. The bootstrap method provides a powerful alternative to arguably more beautiful and CPU-friendly analytic procedures, thus, is a wonderful addition to the hydrologists' arsenal of statistical inference methods. Yet, the apparent simplicity of bootstrapping may fool users into thinking that no important assumptions are being made in its application. These assumptions relate to the independence of the samples and the sample size. Certainly, bootstrapping is not recommended for small training records comprised of only a few observations. Then the resampled records may not be representative of the underlying data generating process and this will corrupt the standard errors and/or confidence intervals of the variables of interest. Furthermore, our experience suggests that it is not particularly easy to preserve higher-order moments (skew and/or kurtosis) of the training data record and/or characterize well the periodicity and persistence of dynamic systems. These are known problems with resampling in general and different implementations of the bootstrap method may be found in the literature to minimize bias and estimation errors. Furthermore, the large computational requirements of the bootstrap method complicate its application to highly-parameterized and/or CPU-intensive models.

3.3. Diagnostic regression

The bootstrap method serves as principal foundation of our methodology for constructing confidence regions and/or intervals of informal goodness-of-fit metrics in regression analysis. This includes metrics such as the NSE and KG efficiency and hydrologic signatures within the context of model diagnostics. We coin this field *diagnostic regression*, not to be confused with regression diagnostics which equal procedures and techniques designed to verify statistical assumptions and model validity in linear regression (Everitt and Skrondal, 2010). Such diagnostic checks are also used in Bayesian inference to ascertain that the residuals satisfy assumptions made by the likelihood function (Schoups and Vrugt, 2010). Thus, in diagnostic regression we model the relationship between dependent and independent variables through application of informal goodness-of-fit metrics and present empirical estimates of the confidence and/or prediction limits of variables of interest. This adds a new member to the large family of commonly used regression techniques such as lasso, logistic, ordinal, polynomial, ridge, support vector, stepwise and quantile regression and offers a common creative license and receptacle for the growing collection of heuristic and/or applied model-data synthesis methods.

Before we move on to the application of diagnostic regression we provide one general remark. We use the wording *empirical confidence intervals* to emphasize the heuristic nature of the confidence intervals obtained from diagnostic regression with metrics such as the KG efficiency. In principle, one could use the label empirical to characterize any sampling-based estimates of the confidence regions and/or intervals. But the so-obtained estimates of the confidence and prediction intervals from the DREAM algorithm are an approximation of the true uncertainty as defined by the residual assumptions.

3.3.1. Application: linear function

We revisit our first case study of Section (3.1) and use the KG efficiency to determine the optimal values of the slope, a, and intercept, b, and their empirical confidence regions and intervals. Bootstrapping involves repeated maximization of the KG efficiency of Eq. (14) for many different realizations of the training data record. Fig. 4 presents the

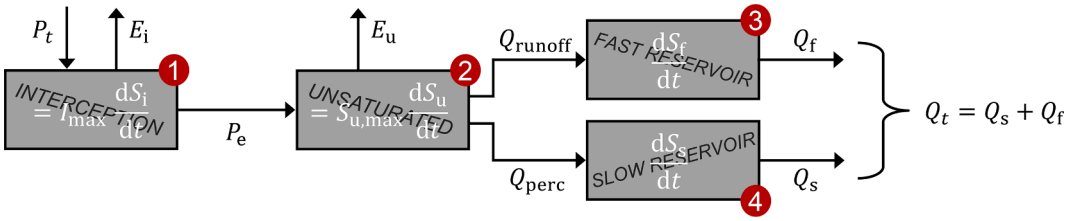

Fig. 5. Schematic illustration of the hmodel after Schoups et al. (2010). Grey boxes, labeled in red, correspond to fictitious control volumes of the watershed which govern the rainfall-runoff transformation. Arrows portray the fluxes into and out of the compartments, including precipitation, P_t, interception evaporation, E_i, surface runoff, Q_{runoff} and percolation, Q_{perc}. The four state variables are simulated using a mass-conservative second-order integration method with adaptive time step.

Table 1
Description of the hmodel parameters, including their symbols, units, lower and upper bounds.

Parameter	Symbol	Units	Min.	Max.
Maximum interception	I_{max}	mm	0	10
Soil water storage capacity	S_{max}	mm	10	1000
Maximum percolation rate	Q_{max}	mm/d	0	100
Evaporation parameter	α_e	–	0	100
Runoff parameter	α_f	–	−10	10
Time constant, fast reservoir	K_f	d	0	10
Time constant, slow reservoir	K_s	d	0	500

confidence regions derived from (a) the analytic expression of Eq. (27) and (b) the bootstrap method using the KG efficiency. The left graph is a copy of Fig. 1a and is used for benchmark purposes.

We observe a close match in the optimum value, $\boldsymbol{\theta}^*$, of the slope and intercept derived from diagnostic regression with the KG efficiency (red cross) and the least squares solution, $\widehat{\boldsymbol{\theta}}$, of Eq. (19). Furthermore, the bivariate confidence regions of the KG efficiency show a strong resemblance with their exact least squares counterparts of Eq. (27). The confidence regions of the KG efficiency center on the optimum (a, b) values, $\boldsymbol{\theta}^*$, and appear well described by ellipses. The major and minor axes of the ellipses of the KG efficiency match quite closely those derived from the information matrix, $\mathscr{I}(\widehat{\boldsymbol{\theta}})$, of the GLS confidence regions, but exhibit an enlarged angle from the horizontal (slope) axis. Furthermore, the ellipses that make up the 95 and 99 confidence regions of a and b appear discontinuous in the area immediately above and below the optimum KG solution (red cross). Indeed, at these critical levels the bootstrap samples provide only a somewhat spotty characterization of the bivariate parameter uncertainty. It is not particularly clear what causes this apparent deficiency. Certainly, we used a large enough sample size.

We do not apply the KG efficiency to our hydrologic toy model of the second case study but rather focus our attention on a more complex watershed model using measured streamflow data instead.

3.3.2. Application: a conceptual watershed model

Our third and last study illustrates the application of diagnostic regression with the KG efficiency to the hmodel, a parsimonious conceptual watershed model originally developed by Schoups et al. (2010).

We estimate the hmodel parameters and their respective confidence intervals using 14-year long records (Oct. 1, 1994 - Sept. 30, 2008) of daily discharge data from the (a) Leaf River near Collins, MS (USGS 02472000) and (b) Kinchafoonee Creek near Dawson, GA (USGS 02350900). These two medium-sized watersheds exhibit a strong and weak winter regime, respectively, according to the functional classification of Brunner et al. (2020). The hmodel transforms rainfall into runoff at the watershed outlet using an interception, unsaturated zone, fast and slow flow reservoir, respectively, which simulate interception, throughfall, evaporation, surface runoff, percolation, fast streamflow and baseflow (see Fig. 5).

The hmodel structure, processes, control input and numerical solution have been discussed by Schoups et al. (2010), and interested readers are referred to this publication for further details. Table 1 lists the seven hmodel parameters and their corresponding symbols, units and upper and lower bounds. We discard the first five years of the discharge records in our computation of the KG efficiency to reduce sensitivity to state variable initialization.

The empirical description of the uncertainty of the KG efficiency with the bootstrap method requires many different replicates of the measured discharge records of the Leaf River and Kinchafoonee Creek. These replicates should characterize streamflow measurement uncertainty and preserve the statistical properties (e.g. streamflow moments and temporal structure/persistence) and hydrologic characteristics (e.g. catchment summary metrics) of the measured discharge record. The model-free duplication method of Oliveira and Vrugt (2022) satisfies these requirements and, thus, serves our purpose. A brief description of this method is given below, interested readers are referred to Oliveira and Vrugt (2022) for further details.

Per Eq. (16), the entries of the discharge measurement vector may be written as follows

$$\tilde{\mathbf{y}} = \mathscr{H}(\mathbf{t}) + \boldsymbol{\epsilon}, \tag{32}$$

where $\mathscr{H}(t)$ is the data generating process of the true streamflow at time $t \in \mathbb{N}_+$ and the measurement errors, $\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2\ \dots\ \epsilon_n]^\top \overset{\mathscr{D}}{\sim} \mathscr{N}_n(\mathbf{0}, \Sigma_\epsilon)$, are variates drawn from a n-variate normal distribution with zero mean and $n \times n$ measurement error covariance matrix, Σ_ϵ

$$\Sigma_\epsilon = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \begin{bmatrix} \sigma_{\epsilon_1}{}^2 & r_y(1)\sigma_{\epsilon_1}\sigma_{\epsilon_2} & \cdots & r_y(n-1)\sigma_{\epsilon_1}\sigma_{\epsilon_n} \\ r_y(1)\sigma_{\epsilon_1}\sigma_{\epsilon_2} & \sigma_{\epsilon_2}^2 & \cdots & r_y(n-2)\sigma_{\epsilon_2}\sigma_{\epsilon_n} \\ \vdots & \vdots & \ddots & \vdots \\ r_y(n-1)\sigma_{\epsilon_1}\sigma_{\epsilon_n} & r_y(n-2)\sigma_{\epsilon_n}\sigma_{\epsilon_2} & \cdots & \sigma_{\epsilon_n}^2 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & r_y(1) & \cdots & r_y(n-1) \\ r_y(1) & 1 & \cdots & r_y(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_y(n-1) & r_y(n-2) & \cdots & 1 \end{bmatrix} \odot \left(\begin{bmatrix} \sigma_{\epsilon_1} \\ \sigma_{\epsilon_2} \\ \vdots \\ \sigma_{\epsilon_n} \end{bmatrix} [\sigma_{\epsilon_1}\ \sigma_{\epsilon_2} \cdots \sigma_{\epsilon_n}] \right) = \mathbf{R}_y \odot \left(\boldsymbol{\sigma}_\epsilon \boldsymbol{\sigma}_\epsilon^\top \right) \tag{33}$$

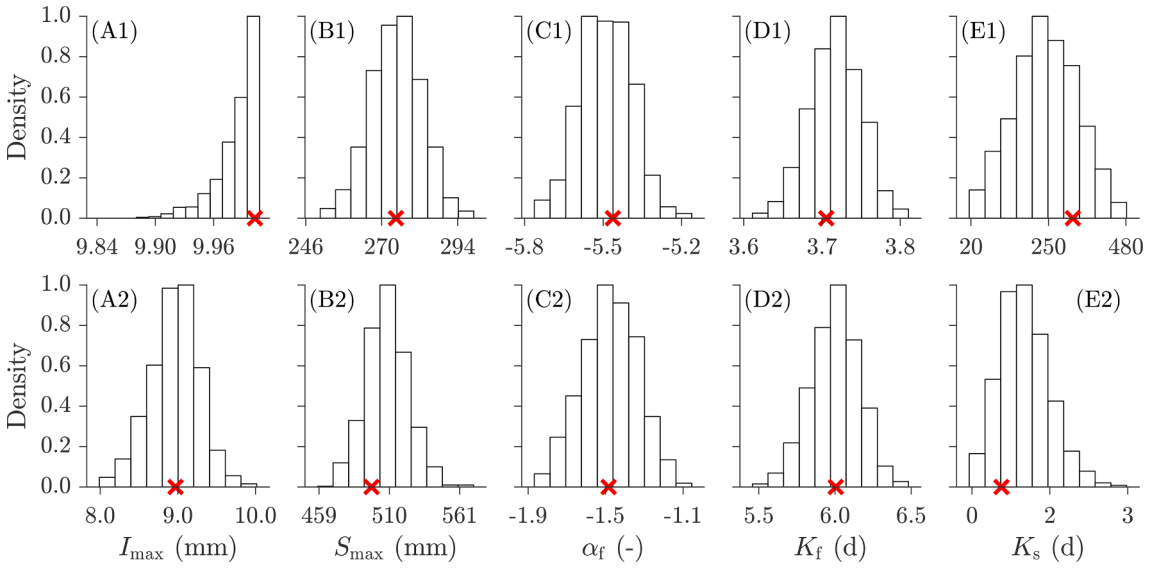

Fig. 6. Histograms of the optimized hmodel parameters for the thousand replicates of the Leaf River (top panel) and Kinchafoonee Creek (bottom panel) watersheds including (a) $S_{\max}$ (mm), (b) $Q_{\max}$ (mm), (c) α_f (-), (d) K_f and (e) K_s. The relative frequencies on the y-axis are normalized to yield a common empirical density between 0 and 1 for all parameters. The red crosses correspond to the optimized hmodel parameter values, $\boldsymbol{\theta}^*$, of the measured discharge records.

where $\mathbf{R}_{\tilde{y}}$ denotes the $n \times n$ correlation matrix of the measurement errors, $r_{\tilde{y}}(\tau)$, is the correlation function of the measurement errors, $\boldsymbol{\sigma}_\epsilon = [\sigma_{\epsilon_1}\ \sigma_{\epsilon_2}\ \dots\ \sigma_{\epsilon_n}]^\top$, signifies the n-vector of measurement error standard deviations, $\odot$, is the Hadamard or Schur product and $\tau \in (1,2,\dots,n)$. The correlation matrix, $\mathbf{R}_{\tilde{y}}$, can be derived from the sample autocorrelation function (ACF) of the measured discharge data, $\tilde{\mathbf{y}} = [\tilde{y}_1\ \tilde{y}_2\ \dots\ \tilde{y}_n]^\top$. The sample autocorrelation, $\hat{r}_{\tilde{y}}(\tau)$, for two streamflow observations, $\tilde{y}_i$ and $\tilde{y}_j$, a distance (time), $\tau = |i-j|$, apart may be computed using

$$\hat{r}_{\tilde{y}}(\tau) = \frac{\mathrm{Cov}[\tilde{y}_i, \tilde{y}_j]}{\mathrm{Var}[\tilde{y}_i]} = \frac{\sum_{i=\tau+1}^{n} \left(\tilde{y}_i - m_{\tilde{y}}\right)\left(\tilde{y}_{i-\tau} - m_{\tilde{y}}\right)}{\sum_{i=\tau+1}^{n} \left(\tilde{y}_i - m_{\tilde{y}}\right)^2}, \tag{34}$$

where $m_{\tilde{y}} = \frac{1}{n}\sum_{t=1}^{n} \tilde{y}_t$ (mm/d) denotes the mean of the n-record of streamflow observations. The entries of the $n \times 1$ vector of measurement error variances, $\boldsymbol{\sigma}_\epsilon^2$, are computed from hourly discharge observations using the procedure described in Oliveira and Vrugt (2022). Specifically, the t^{th}-entry, $s_{\epsilon_t}^2$, of s_ϵ^2, is computed as follows (Oliveira and Vrugt, 2022)

$$s_{\epsilon_t}^2 = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}\left(\tilde{y}_{ih} - \frac{1}{m}\sum_{j=1}^{m}\tilde{y}_{jh}\right)^2 + \frac{1}{m}\frac{1}{m}\sum_{i=1}^{m}(a_h\tilde{y}_{ih} + b_h)^2, \tag{35}$$

where the $\tilde{y}_{ih}$'s (mm/d) are the $m = 24$ hourly discharge observations, $i = (1,2,\dots,m)$, of the t^{th} day of the daily streamflow record, $\tilde{\mathbf{y}}$, and the coefficients, a_h (-) and b_h (mm/d) signify the slope and intercept of the hourly discharge measurement error function, respectively. This linear function turns the hourly discharge measurements into estimates of the measurement error standard deviation and is derived from nonparametric differencing using the estimator of Vrugt et al. (2005). Eq. (35) can be rewritten as follows

$$s_{\epsilon_t}^2 = \frac{1}{m}s_{y_h}^2 + \frac{1}{m^2}\sum_{i=1}^{m} s_{\epsilon_h}^2 \tag{36}$$

where the first term, $s_{y_h}^2$ (mm^2/d^2), measures the spread of the hourly discharge observations (= measurement uncertainty) and the second term, $\frac{1}{m}s_{\epsilon_h}^2$ (mm^2/d^2), accounts for their respective measurement error variances.

Replicate discharge records, $\tilde{\mathbf{y}}_r = \tilde{\mathbf{y}} + \boldsymbol{\epsilon}$, are now created by perturbing the measured streamflow time series, $\tilde{\mathbf{y}}$, with measurement errors, $\boldsymbol{\epsilon}$, drawn from the n-variate normal distribution, $\boldsymbol{\epsilon} \overset{\mathscr{D}}{\sim} \mathcal{N}_n(\mathbf{0}, \Sigma_\epsilon)$, with zero mean and covariance matrix, Σ_ϵ, of Eq. (33). To do so efficiently we write instead, $\tilde{\mathbf{y}}_r = \tilde{\mathbf{y}} + \mathbf{L}\boldsymbol{\nu}$, where $\boldsymbol{\nu} \in \mathbb{R}^n$ is a $n \times 1$ vector of independent standard normal variates and the $n \times n$ lower triangular matrix $\mathbf{L}$ is derived from Cholesky decomposition of the symmetric positive-definite matrix $\Sigma_\epsilon = \mathbf{L}\mathbf{L}^\top$. Prior to computation of the correlation matrix, $\mathbf{R}_{\tilde{y}}$, the discharge measurements are replaced by their respective variates of a standard normal distribution. This transformation promotes hydrologic characterization by suppressing sudden bumps and discharge fluctuations during long recession periods. Note that if Σ_ϵ is written as product of a constant, σ_ϵ^2, and a $n \times n$ matrix, $\mathbf{V}$, then, $\mathbf{L} = \sqrt{\sigma_\epsilon}^2\mathrm{chol}(\mathbf{V})$. The use of the discharge sample ACF in the measurement error covariance matrix, Σ_ϵ, introduces serial correlation among the ϵ_t's. This is a necessary means to preserving the smoothness, statistical and hydrologic properties of the measured discharge record (Oliveira and Vrugt, 2022). The autocorrelation avoids overly bumpy replicate records that result from the use of the measured discharge time series rather than the underlying data generating process in Eq. (32).

Fig. 6 presents histograms of a representative group of five hmodel parameters (a) $S_{\max}$, (b) $Q_{\max}$, (c) α_f, (d) K_f and (e) K_s derived from the bootstrap method using the $N = 1,000$ replicates of the measured daily discharge records of the Leaf River (top panel) and Kinchafoonee Creek (bottom panel). For each replicate record, we optimized the seven hmodel parameters by maximization of the KG efficiency using the shuffled complex evolution (SCE-UA) algorithm of Duan et al. (1992).

The marginal distributions summarize the effect of the discharge measurement errors on the inferred hmodel parameter values. The modes of the histograms coincide quite well with their optimized values, $\boldsymbol{\theta}^*$, of the measured discharge records (red crosses). The frequency distributions of the hmodel parameters appear well defined by calibration against the KG efficiency. The parameters exhibit a relatively small dispersion, appear symmetric around their mean and are well described by a Gaussian distribution. Exceptions to this are the histograms of parameters $I_{\max}$ (Fig. 3a1) which is truncated by its upper bound and K_s (Fig. 3e2) which has a positive skew, and, thus, tail to the right. As a result, the marginal distributions of these two parameters do not center on their values derived from the measured discharge record but are

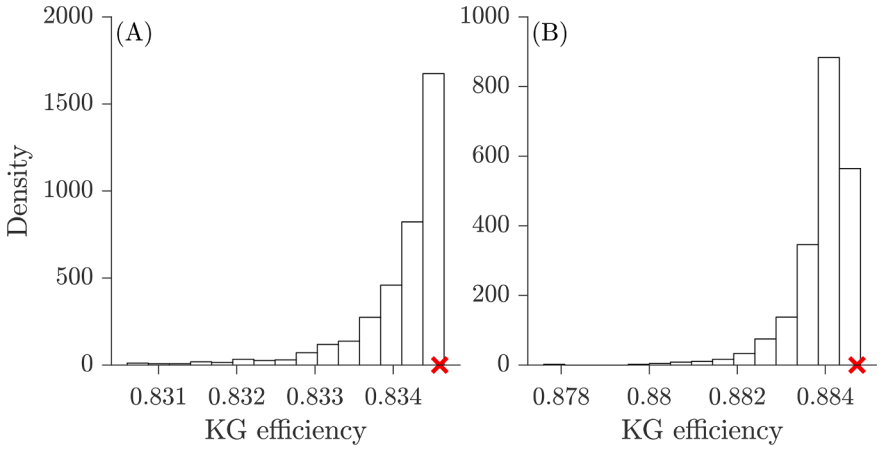

Fig. 7. Histogram of the KG efficiency derived from repeated optimization of the hmodel parameters using the replicates of the discharge record of the (a) Leaf River and (b) Kinchafoonee Creek. The maximized KG efficiencies of the measured discharge data are separately indicated with a red cross.

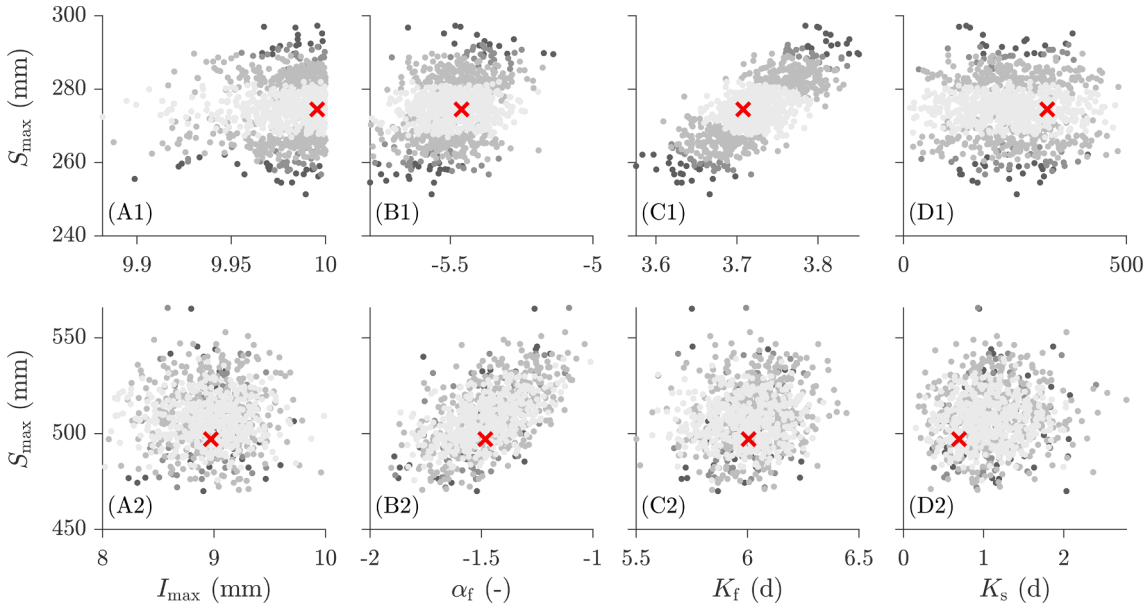

Fig. 8. Bivariate scatter plots of selected parameter pairs including (a) (I_{max}, S_{max}), (b) (α_f, S_{max}), (c) (K_f, S_{max}) and (d) (K_s, S_{max}) for the Leaf River (top panel) and Kinchafoonee Creek (bottom panel) watersheds. The confidence regions are coded with a gray color scheme using $\gamma = 0.50$ (light), $\gamma = 0.90$ (light-medium), $\gamma = 0.95$ (medium) and $\gamma = 0.99$ (dark), respectively. The red crosses signify the hmodel parameter values that maximize the KG efficiency of the measured discharge records.

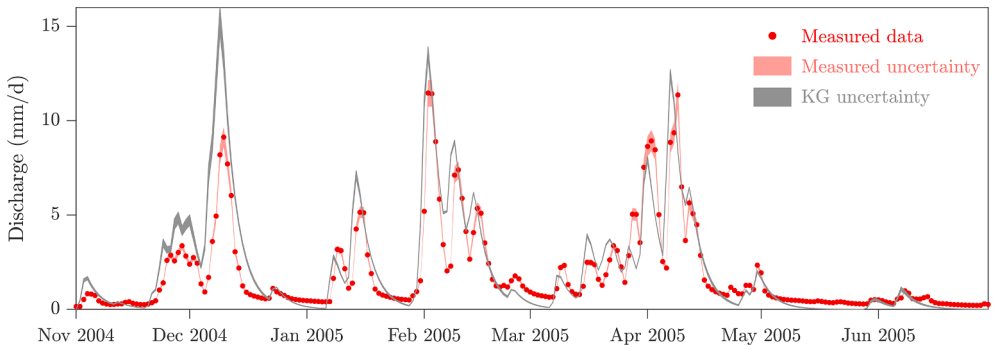

Fig. 9. Observed (red dots) and 99% hmodel simulated (gray region) daily discharge time series for a seven month period between Nov. 1, 2004 and June 30, 2005 of the discharge record of the Leaf River basin. The light-red region corresponds to the 99% intervals of the discharge measurement errors.

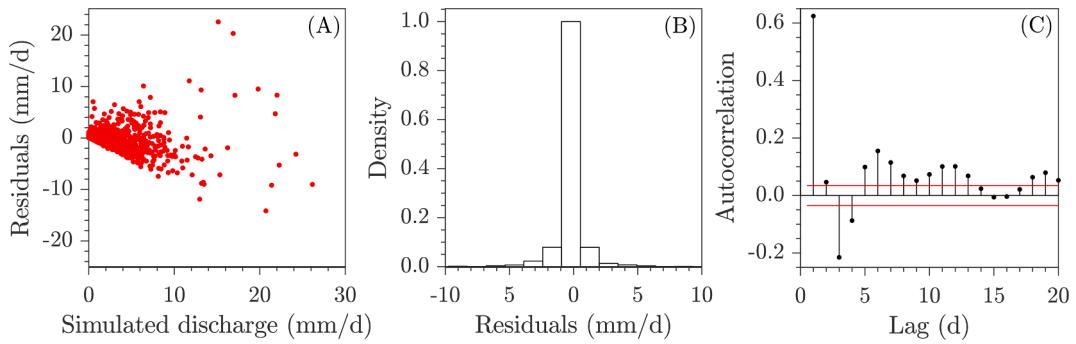

Fig. 10. Diagnostic analysis of the streamflow residuals of the calibrated hmodel using the KG efficiency: (a) residuals as a function of simulated discharge, (b) histogram of residuals, (c) ACF with 95% significance levels (dotted red lines).

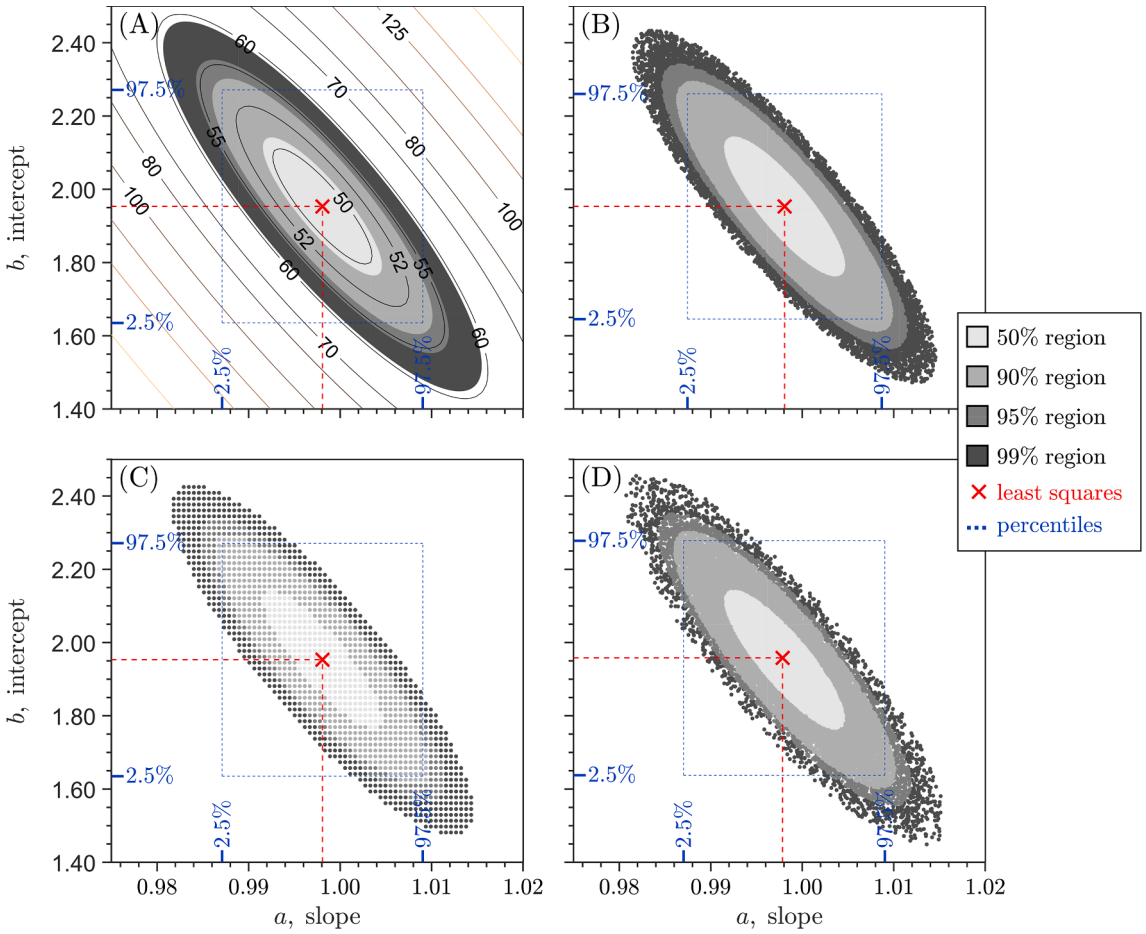

Fig. 11. Confidence regions (gray tints) and 95% confidence intervals (dashed blue lines) for a linear regression function, $f(\boldsymbol{\theta}, t) = at + b$, with homoscedastic measurement errors and $0.97 \leqslant a \leqslant 1.025$ and $1.30 \leqslant b \leqslant 2.60$: (a) generalized least squares using the analytic expressions of Eqs. (27) and (29), (b) resampling of the training data record with the bootstrap method, (c) contouring of the GLS objective function using the tolerable increment of Eq. (C3) and (d) Bayesian inference with the DREAM algorithm using a uniform prior for the slope, a, and intercept, b, and likelihood function of Eq. (C4). The bivariate $100\gamma\%$ confidence regions of the slope, a, and intercept, b, are displayed with a gray color scheme using $\gamma = 0.50$ (light gray), $\gamma = 0.90$ (light-medium gray), $\gamma = 0.95$ (medium gray) and $\gamma = 0.99$ (dark gray), respectively. The top-left graph includes contour lines of the GLS objective function in Eq. (C1). The red cross highlights the minimum of the GLS objective function.

found at the upper and lower end of their distribution. This deviation from normality may become more common for catchments from other hydrologic regimes, for example with snow melt or extended dry periods.

Next, Fig. 7, presents the marginal distribution of the KG efficiency obtained from evaluating the optimized hmodel parameters of the thousand replicate records for the measured discharge data of the (a) Leaf River (top panel) and (b) Kinchafoonee Creek. The optimized KG values of the measured discharge data records are separately indicated with a blue cross.

As the KG efficiency is a metric that has to be maximized, its frequency distribution is truncated at the upper bound by its maximized value for the measured discharge record and the probability mass of the bootstrap samples is dispersed in a tail to the left. Quite remarkably, the 50, 90, 95 and 99% percentiles of the distributions of the KG efficiency have a similar distance to the maximum KG efficiency. This amounts to

0.0032, 0.0018, 0.0013, and 0.005, respectively, listed in order of the critical values. This finding may support a more formal probabilistic description of the KG efficiency in a fashion similar to the tolerable increments for the GLS estimator discussed in Appendix C. Indeed, for the GLS estimator, the tolerable increment follows an inverse chi-square cumulative distribution function with p degrees of freedom, hence, $\Delta\text{GLS}(\gamma) = F^{-1}_{\chi^2}(p_\gamma|p)$, where $p_\gamma = \gamma$. Thus, the tolerable increment of the GLS only depends on parameter dimensionality, not on data length and the nature of the measurement errors (residuals). We have investigated this thread for the KG efficiency with the linear regression function of Section 3.1 using different lengths of the training data record and homoscedastic, heteroscedastic and/or correlated measurement errors. Our preliminary results (not shown) have demonstrated that the tolerable reduction of the KG efficiency, ΔKG, depends on the magnitude and nature of the measurement errors. In other words, the KG efficiency does not admit a simple probabilistic description of its confidence regions and/or intervals. This leaves as our only option the bootstrap method using replicates of the discharge record.

To provide insights into the bootstrap samples, please consider Fig. 8 which presents bivariate scatter plots of (a) $(I_{\max}, S_{\max})$, (b) $(\alpha_f, S_{\max})$, (c) $(K_f, S_{\max})$ and (d) $(K_s, S_{\max})$ using the replicates of the discharge record of the Leaf River (top panel) and Kinchafoonee Creek (bottom panel) watersheds. The red crosses portray the optimal values of the hmodel parameters, $\boldsymbol{\theta}^*$, using diagnostic regression for the measured discharge records. The bootstrap samples are coded in different gray tints based on their confidence levels, $\gamma = 0.5$ (light gray), $\gamma = 0.90$ (light-medium gray), $\gamma = 0.95$ (medium gray) and $\gamma = 0.99$ (dark gray).

The bootstrap samples populate only a small part of the prior parameter space. The dotty plots are well described by concentric circles and/or elongated ellipses which center on the midpoint of the dotty plots and envelope the optimum solution (red cross) of the hmodel parameters. The exception to this is the (a1) $(I_{\max}, S_{\max})$ parameter pair whose bivariate distribution cloud is truncated by the upper bound of $I_{\max}$. The density of the points decreases away from the midpoint of the point cloud. We do not witness any mutual relationships between the pairs of plotted parameters, with exception of the (b2) $(\alpha_f, S_{\max})$ and (c1) $(K_f, S_{\max})$ parameter pairs, which exhibit a positive linear correlation. The bivariate confidence regions of the hmodel parameters are sharply delineated for the Leaf River data record and organized in long elongated ellipses around the midpoint of the bootstrap samples. For the Kinchafoonee Creek, on the contrary, the boundaries of the joint parameter confidence regions of the hmodel parameters are poorly defined. The confidence regions mix and overlap, an effect that is particularly visible towards the outer perimeter of the point clouds of the hmodel parameters pairs. This mixing of the confidence regions is a result of the projection of the $p = 7$-dimensional parameter space onto only two axes. This geometric simplification destroys the underlying multivariate surface of the KG efficiency and this distorted organization may, therefore, lead to a mixing of the confidence regions. The reason so as to why the Leaf River watershed does not suffer this protrusion conveys important information about the geometry of the KG confidence regions in the full parameter space. In short, the confidence regions can only increase (or decrease) monotonically with each marginalized (unplotted) parameter axes for the spatial organization of the confidence regions to be preserved in the two-dimensional projections in the top panel. The p-variate confidence regions for the Kinchafoonee Creek violate the monotonicity requirement for at least one of the marginalized parameter axes, and as a result, the projection introduces mixing so evidently present in the two-dimensional snapshots in the bottom panel. Of course, this mixing will not affect the confidence intervals of the hmodel parameters.

Finally, we must verify the accuracy and/or precision of the KG calibrated hmodel by comparison against the measured discharge record. Fig. 9 presents a time series plot of measured (red dots) and hmodel simulated discharge for the Leaf River watershed using a representative 8-month portion of the historical record. As the Kinchafoonee Creek presents similar findings, we do not visualize these results. The light red area displays the 99% discharge measurement uncertainty, whereas the gray region corresponds to 99% hmodel simulation uncertainty associated with the KG efficiency. We do not display the streamflow simulation of the optimal hmodel parameters, $\boldsymbol{\theta}^*$, of the measured discharge record of the Leaf River. This simulation is contained within the 99% confidence limits of the simulated discharge, mostly at the center of the gray interval.

The hmodel tracks the measured discharge data reasonably well, although a positive bias is observed in the first 50-days of the 7-month period between the middle-end of November and middle of December, 2004. This initial overshoot of the measured discharge data is rectified during subsequent storm events demonstrating an increasingly better match between the simulated and measured hydrographs. Certainly, the baseflow is rather poorly described by the hmodel in the 242-day window. This deficiency is particularly visible in the long recession period at the end of the 7-month record. The large model-data mismatch so visible in the early part of the record may point at an exaggerated state of the hmodel's four control volumes in the period leading up to the storm event in November 2004. Measurement errors of the antecedant basin average rainfall may have corrupted the state variables of the hmodel. One has to be careful, however, in attributing this early mismatch to precipitation errors as performance metrics other than the KG efficiency may improve the overall compliance between the hmodel and the measured discharge data. This then would have an immediate effect on the simulated state of the Leaf River basin preceding the storm event in Nov. 2004. If so desired, data assimilation may be used to refine the simulated state variables and remove excess water from the control volumes if the measured discharge data dictate doing so improves model-data compliance.

The 99% confidence intervals of the simulated hydrograph (gray region) are smallest, on average, at the end of a long recession period and reach a maximum width at peak discharge. This dependence of the width of the hmodel discharge confidence intervals on simulated flow level is a result of the heteroscedastic nature of the discharge measurement errors as evidenced by the measured streamflow record using nonparametric differencing (Vrugt et al., 2005; Oliveira and Vrugt, 2022). The 99% confidence intervals of the hmodel simulated streamflows exceed the discharge measurement uncertainty (red region) but appear rather small compared to the residuals of the optimal hmodel parameters, $\boldsymbol{\theta}^*$. Indeed, the gray region makes up only a small part of the distance between the measured and simulated discharge records. This small width is commensurate with a poor coverage of the discharge confidence intervals of the KG efficiency. Only a handful of streamflow observations contained within the gray region. This is why it is common practice in the context of generalized least squares to use the sample variance of the residuals (see Eq. (24)) in the computation of the 100γ% confidence intervals of the GLS parameters, $\widehat{\boldsymbol{\theta}}$, and simulated output, $\widehat{\mathbf{y}}$, in Eqs. (29) and (30). We can follow a similar approach in diagnostic regression and use the residuals of the optimal hmodel parameters, $\boldsymbol{\theta}^*$, in our computation of the empirical confidence intervals of the KG efficiency. This requires only a minor change to the implementation of the bootstrap method. We must replace our probabilistic description of the measurement errors of the training data record with an analogous description of the model residuals of the maximized KG efficiency of the measured training data record. Inevitably, this will substantially enhance hmodel parameter and simulation uncertainty. The formulation of the n-variate residual distribution is relatively simple for well-behaved residuals with known marginal distribution, constant variance and/or structure that satisfies a simple autoregressive scheme. But skewed and/or lepto- or platykurtic residuals with a nonconstant variance, bias and/or an unusual structure (persistence and/or state dependence) do not necessarily admit a convenient probabilistic description. This limits our ability to draw accurate replicates of the

training data record and complicates uncertainty quantification of model parameter and simulation uncertainty with the bootstrap method in diagnostic regression with the KG efficiency. The generalized plus and universal likelihood functions of Vrugt et al. (2022) will help in distilling a convenient probabilistic description of time series of ideal and non-ideal residuals. But this does not guarantee a perfect characterization.

The time series plot in Fig. 9 certainly makes clear that the discharge residuals exhibit a changing bias and correlation structure with flow level. But as the KG efficiency is an informal goodness-of-fit metric, we cannot unify the actual residual characteristics with prior assumptions made about their probabilistic properties. Nevertheless, Fig. 10 analyzes the (a) magnitude, (b) distribution and (c) ACF of the streamflow residuals of the calibrated hmodel with maximum KG efficiency.

The residuals increase with magnitude of the simulated discharge in a manner that is expected from the knowledge of the measurement errors. The empirical distribution of the discharge residuals appears symmetric, is centered about zero and is much peakier than the normal distribution. Lastly, the residuals exhibit considerable serial correlation at the first few lags. This is expected given the systematic over and/or underprediction so evidently present at high and low flows, respectively, in the time series plot. These findings warrant treatment of residual serial correlation and heteroscedasticity, for example, through the use of generalized least squares. Alternatively, we can resort to Bayesian analysis, specify a generalized likelihood function (Schoups and Vrugt, 2010) and infer the matrix $\mathbf{V}$ along with σ_ϵ^2 simultaneously with the hmodel parameters using MCMC simulation with the DREAM algorithm. This approach lets the data speak for itself and provides samples from the posterior distribution which can be presented as confidence regions.

4. Conclusions

Informal quality-of-fit measures such as the KG efficiency are not borne out of testable hypotheses with respect to the probabilistic properties of the residuals. This has profound consequences. We cannot verify a posteriori whether assumptions of the KG estimator have been satisfied. And, more importantly from the perspective of this paper, the uncertainty of the KG efficiency is not defined. This begs the question, how we should compute confidence and prediction limits on current and/or future model responses if we do not know which marginal distribution to expect of the residuals of the KG efficiency?.

To move beyond the status quo, this paper has presented a simple framework for determining empirical confidence intervals of the KG efficiency. Our method relates the distribution of the KG efficiency to the measurement errors of the calibration data. Parameter and simulation uncertainty may then be quantified with the bootstrap method using replicates of the data record.

The first two case studies served as demonstration of the bootstrap method for statistical inference of parameter uncertainty to those unfamiliar with this methodology. We showed that the bootstrap method yields the exact same parameter confidence regions and intervals as generalized least squares within the context of linear regression and Bayesian analysis coupled with MCMC simulation within the context of nonlinear regression.

After this proof of concept, we turned our attention to the KG efficiency and used this informal goodness-of-fit metric within the context of diagnostic regression to determine optimal parameter values and their associated uncertainty of a simple linear regression function with slope and intercept and the 7-parameter hmodel of Schoups et al. (2010) using measured discharge data of two contrasting watersheds.

The empirical confidence regions and intervals of the slope and intercept derived from diagnostic regression using the KG efficiency were in close agreement with their exact counterparts obtained from generalized least squares. The ellipses of the KG efficiency exhibited an enlarged angle from the horizontal axis.

The application of diagnostic regression with the KG efficiency to the hmodel showed that its parameters were well described by a normal distribution with relatively small dispersion and/or skew, and, possibly, truncated by the prior distribution. The modes of the marginal parameter distributions coincided quite well with their optimized values derived from the measured discharge records of the Leaf River and Kinchafoonee Creek. The distribution of the KG efficiency is a complex function of data length and the magnitude, distribution and structure of the discharge measurement errors. This prohibits a simple closed-form description of the empirical confidence regions and/or intervals of the KG efficiency defined herein. This leaves as only option the bootstrap method to quantify model parameter and predictive uncertainty of the KG efficiency within the context of diagnostic regression.

Data and Software Availability

The data, models and other software are available upon request from the corresponding author, `jasper@uci.edu`, and can be downloaded from https://github.com/jaspervrugt/KGefficiency.

CRediT authorship contribution statement

Jasper A. Vrugt: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Debora Y. de Oliveira:** Software, Validation, Formal analysis, Data curation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We greatly appreciate the constructive comments of the AE and two reviewers that have led to an improved manuscript. The first author acknowledges interactions with Dr. Yan Liu on the mathematical underpinning of the KG estimator. The second author gratefully acknowledges the financial support received from the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), Grant No. 88881.174456/2018–01. The CAMELS data set is described in Newman (2015) and can be downloaded from https://dx.doi.org/10.5065/D6MW2F4D. The hourly streamflow data of Gauch et al. (2020) are available at https://doi.org/10.5281/zenodo.4072700.

Appendix A. Decomposition of the mean squared residual

The original derivation of the mean squared residual (MSR) by Gupta et al. (2009) assumes knowledge of the population variances of the $n \times 1$ records of measured, $\tilde{\mathbf{y}} = [\tilde{y}_1\ \tilde{y}_2\ \dots\ \tilde{y}_n]^\top$, and simulated, $\mathbf{y} = [y_1\ y_2\ \dots\ y_n]^\top$, data.

The MSR is equal to

$$\mathrm{MSR} = \frac{1}{n}\sum_{t=1}^{n}\left(\widetilde{y}_t - y_t(\boldsymbol{\theta})\right)^2 \tag{A1}$$

and may be decomposed in different terms as follows

$$\mathrm{MSR} = \frac{1}{n}\sum_{t=1}^{n}\left(\widetilde{y}_t - y_t(\boldsymbol{\theta})\right)\left(\widetilde{y}_t - y_t(\boldsymbol{\theta})\right) = \frac{1}{n}\sum_{t=1}^{n}\widetilde{y}_t^2 + \frac{1}{n}\sum_{t=1}^{n}y_t(\boldsymbol{\theta})^2 - \frac{2}{n}\sum_{t=1}^{n}\widetilde{y}_t y_t(\boldsymbol{\theta}), \tag{A2}$$

We can now make use of the following well-known identities to rewrite the above expression

$$\sum_{i=1}^{n}x_i^2 = \left(n-1\right)s_x^2 + nm_x^2 \tag{A3a}$$

$$\sum_{i=1}^{n}x_i y_i = \left(n-1\right)r_{xy}s_x s_y + nm_x m_y, \tag{A3b}$$

where m_x is the mean of the n-record of x values

$$m_x = \frac{1}{n}\sum_{i=1}^{n}x_i, \tag{A4}$$

the variable s_x^2 denotes its associated variance

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - m_x\right)^2. \tag{A5}$$

and r_{xy} signifies the sample correlation coefficient of the (x_i, y_i) data pairs, $i = (1, 2, \ldots, n)$

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)\left(y_i - m_y\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{y}_i - m_{\widetilde{y}}\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - m_y\right)^2}}. \tag{A6}$$

If we substitute the identities of Eqs. (A3a) and (A3b) into Eq. (A2) we yield

$$\begin{aligned}\mathrm{MSR} &= \frac{1}{n}\left(\left(n-1\right)s_{\widetilde{y}}^2 + nm_{\widetilde{y}}^2\right) + \frac{1}{n}\left(\left(n-1\right)s_y^2 + nm_y^2\right) - \frac{2}{n}\left(\left(n-1\right)r_{\widetilde{y}y}s_{\widetilde{y}}s_y + nm_{\widetilde{y}}m_y\right) = \left(\frac{n-1}{n}\right)s_{\widetilde{y}}^2 + m_{\widetilde{y}}^2 + \left(\frac{n-1}{n}\right)s_y^2 + m_y^2 - \left(\frac{2n-2}{n}\right)r_{\widetilde{y}y}s_{\widetilde{y}}s_y - 2m_{\widetilde{y}}m_y \\ &= \left(\frac{n-1}{n}\right)\left(s_{\widetilde{y}}^2 + s_y^2\right) + \left(m_{\widetilde{y}} - m_y\right)^2 - \left(\frac{2n-2}{n}\right)r_{\widetilde{y}y}s_{\widetilde{y}}s_y\end{aligned} \tag{A7}$$

Now, as, $a^2 + b^2 = (a-b)^2 + 2ab$, we can rewrite the first term to read

$$\begin{aligned}\mathrm{MSR} &= \left(\frac{n-1}{n}\right)\left(\left(s_{\widetilde{y}} - s_y\right)^2 + 2s_{\widetilde{y}}s_y\right) + \left(m_{\widetilde{y}} - m_y\right)^2 - \left(\frac{2n-2}{n}\right)r_{\widetilde{y}y}s_{\widetilde{y}}s_y = \left(\frac{n-1}{n}\right)\left(s_{\widetilde{y}} - s_y\right)^2 + \left(m_{\widetilde{y}} - m_y\right)^2 + \left(\frac{2n-2}{n}\right)s_{\widetilde{y}}s_y - \left(\frac{2n-2}{n}\right)r_{\widetilde{y}y}s_{\widetilde{y}}s_y \\ &= \left(\frac{n-1}{n}\right)\left(s_{\widetilde{y}} - s_y\right)^2 + \left(m_{\widetilde{y}} - m_y\right)^2 + \left(\frac{2n-2}{n}\right)s_{\widetilde{y}}s_y\left(1 - r_{\widetilde{y}y}\right)\end{aligned} \tag{A8}$$

We can now reorganize the above expression and rearrange the terms in similar order as Eq. (7) to yield

$$\mathrm{MSR} = \left(\frac{2n-2}{n}\right)s_{\widetilde{y}}s_y\left(1 - r_{\widetilde{y}y}\right) + \left(\frac{n-1}{n}\right)\left(s_{\widetilde{y}} - s_y\right)^2 + \left(m_{\widetilde{y}} - m_y\right)^2 \tag{A9}$$

This concludes our derivation.

Appendix B. Decomposition of the coefficient of determination

According to Eq. (6) the coefficient of determination, R^2, satisfies the following equality

$$R^2 = 1 - \frac{n\mathrm{MSR}}{\left(n-1\right)s_{\widetilde{y}}^2} = \mathrm{NSE}, \tag{B1}$$

where NSE is the infamous Nash–Sutcliffe efficiency. We can reformulate the above expression by substituting for the mean squared residual (MSR) Eq. (A9) to yield

$$R^2 = 1 - \frac{\left(2n-2\right)s_{\tilde{y}}s_y\left(1-r\right) + \left(n-1\right)\left(s_{\tilde{y}} - s_y\right)^2 + n\left(m_{\tilde{y}} - m_y\right)^2}{\left(n-1\right)s_y^2} = \frac{\left(n-1\right)s_{\tilde{y}}^2 - \left(2n-2\right)s_{\tilde{y}}s_y\left(1-r\right) - \left(n-1\right)\left(s_{\tilde{y}} - s_y\right)^2 - n\left(m_{\tilde{y}} - m_y\right)^2}{\left(n-1\right)s_y^2}$$
$$= \frac{\left(n-1\right)s_{\tilde{y}}^2 + \left(2n-2\right)s_{\tilde{y}}s_y\left(r-1\right) - \left(n-1\right)\left(s_{\tilde{y}}^2 + s_y^2 - 2s_{\tilde{y}}^2 s_y\right) - n\left(m_{\tilde{y}} - m_y\right)^2}{\left(n-1\right)s_y^2}$$
$$= \frac{\left(n-1\right)s_{\tilde{y}}^2 + \left(2n-2\right)s_{\tilde{y}}s_y\left(r-1\right) - \left(n-1\right)s_{\tilde{y}}^2 - \left(n-1\right)s_y^2 + 2\left(n-1\right)s_{\tilde{y}}^2 s_y - n\left(m_{\tilde{y}} - m_y\right)^2}{\left(n-1\right)s_y^2} = \frac{\left(2n-2\right)s_{\tilde{y}}s_y r - \left(n-1\right)s_y^2 - n\left(m_{\tilde{y}} - m_y\right)^2}{\left(n-1\right)s_y^2}. \tag{B2}$$

This leaves us with the following expression for the R^2 and, thus, NSE

$$R^2 = 2\left(\frac{s_y}{s_{\tilde{y}}}\right)r - \left(\frac{s_y}{s_{\tilde{y}}}\right)^2 - \left(\frac{n}{n-1}\right)\left(\frac{m_{\tilde{y}} - m_y}{s_{\tilde{y}}}\right)^2 \tag{B3}$$

This concludes the derivation.

Appendix C. Description of multivariate parameter uncertainty

The confidence regions described by the expression in Eq. (27) can also be inferred by other means. In this Appendix we consider two other approaches besides the bootstrap method described in the main text. The first of these alternative methods uses contouring of the GLS objective function.

Consider the GLS objective function in Eq. (17)

$$F_{\mathrm{GLS}}(\boldsymbol{\theta}) = \mathbf{e}(\boldsymbol{\theta})^\top \Sigma_\epsilon^{-1}\mathbf{e}(\boldsymbol{\theta}), \tag{C1}$$

which may also be written as follows

$$F_{\mathrm{GLS}}(\boldsymbol{\theta}) = (\mathbf{W}\mathbf{e}(\boldsymbol{\theta}))^\top(\mathbf{W}\mathbf{e}(\boldsymbol{\theta})) = \boldsymbol{\varepsilon}(\boldsymbol{\theta})^\top\boldsymbol{\varepsilon}(\boldsymbol{\theta}), \tag{C2}$$

where $\mathbf{W} = \Sigma_\epsilon^{-\frac{1}{2}}$ signifies the $n \times n$ weight matrix and, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = [\varepsilon_1(\boldsymbol{\theta})\ \varepsilon_2(\boldsymbol{\theta}) \ldots \varepsilon_n(\boldsymbol{\theta})]^\top$, denotes the $n \times 1$ vector of homogenized and/or decorrelated residuals. Now, we expect, that if $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, then the n squared entries of $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$, should, on average, have a value of unity. As a result, $F_{\mathrm{GLS}}(\widehat{\boldsymbol{\theta}})$ should follow a chi-square distribution with $n-p$ degrees of freedom, hence, $F_{\mathrm{GLS}}(\widehat{\boldsymbol{\theta}}) \sim \chi^2_{n-p}$, with expected value, $\mathbb{E}(\boldsymbol{\varepsilon}(\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\varepsilon}(\widehat{\boldsymbol{\theta}})) = n-p$. Any deviation of the parameters from $\widehat{\boldsymbol{\theta}}$, will increase the value of $F_{\mathrm{GLS}}(\widehat{\boldsymbol{\theta}})$ from its expected minimum of $n-p$. The larger this increment, the lesser the support for the parameter values, $\boldsymbol{\theta}$, by the training data, $\tilde{\mathbf{y}}$. The tolerable increment, ΔF_{GLS}, from $n-p$, for a desired confidence level, γ, equals (Press et al., 1992)

$$\Delta F_{\mathrm{GLS}}\left(\gamma\right) = F_{\chi^2}^{-1}\left(p_\gamma \middle| p\right) \tag{C3}$$

where $F_{\chi^2}^{-1}(p_\gamma|p)$ signifies the inverse of the chi-square cumulative distribution function (cdf) with p degrees of freedom at the critical value, $p_\gamma = \gamma$. Now, we discretize the parameter space in uniform intervals and evaluate the objective function, $F_{\mathrm{GLS}}(\boldsymbol{\theta})$, at each grid point. All points, $\boldsymbol{\theta}$, with $F_{\mathrm{GLS}}(\boldsymbol{\theta}) \leqslant F_{\mathrm{GLS}}(\widehat{\boldsymbol{\theta}}) + \Delta F_{\mathrm{GLS}}(\gamma)$ will make up the 100γ% confidence region of the parameters (see Fig. 11c).

As second approach we consider Monte Carlo simulation. This approach necessitates the use of a prior distribution and a likelihood function. To be commensurate with the GLS objective function, we must specify a uninformative prior distribution for the slope and intercept, a and b, and use the following formulation of the log-likelihood function, $\mathscr{L}(\boldsymbol{\theta}, \widehat{\sigma}_\epsilon^2|\tilde{\mathbf{y}})$

$$\mathscr{L}\left(\boldsymbol{\theta}, \widehat{\sigma}_\epsilon^2 \middle| \tilde{\mathbf{y}}, \mathbf{V}\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\widehat{\sigma}_\epsilon^2\mathbf{V}\right|\right) - \frac{1}{2}\mathbf{e}(\boldsymbol{\theta})^\top\left(\widehat{\sigma}_\epsilon{}^2\mathbf{V}\right)^{-1}\mathbf{e}(\boldsymbol{\theta})^\top \tag{C4}$$

where $|\cdot|$ signifies the determinant operator and, $\widehat{\sigma}_\epsilon^2$, is the estimate of the population variance of the measurement errors, ϵ. To be comparable with our GLS implementation, we must infer its value jointly with those of the coefficients, a and b, of the linear regression function. We use Markov chain Monte Carlo simulation with the DREAM algorithm (Vrugt et al., 2009) to determine the trivariate posterior distribution of $\boldsymbol{\theta} = [a\ b]^\top$ and $\widehat{\sigma}_\epsilon^2$. Fig. 11d presents the bivariate distribution of the slope and intercept. As expected (not shown), the marginal distribution of $\widehat{\sigma}_\epsilon^2$, follows a scaled chi-square distribution with $n-p$ degrees of freedom (see Eq. (25)).

References

Aitken, A.C., 1936. Iv.-on least squares and linear combination of observations. Proc. R. Soc. Edinb. 55, 42–48. https://doi.org/10.1017/S0370164600014346.

Ammann, L., Fenicia, F., Reichert, P., 2019. A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. Hydrol. Earth Syst. Sci. 23 (4), 2147–2172. https://doi.org/10.5194/hess-23-2147-2019.

Anderson, T., Darling, D.A., 1952. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. Ann. Math. Stat. 23, 193–212. https://doi.org/10.1214/aoms/1177729437.

Barber, C., Lamontagne, J., Vogel, R.M., 2019. Improved estimators of correlation and r2 for skewed hydrologic data. Hydrol. Sci. J. 65 (1), 87–101. https://doi.org/10.1080/02626667.2019.1686639.

Bates, B.C., Campbell, E.P., 2001. A markov chain monte carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour. Res. 37 (4), 937–947. https://doi.org/10.1029/2000WR900363.
Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.
Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305.
Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. J. Hydrol. 249 (1), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8.
Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36 (12), 3663–3674. https://doi.org/10.1029/2000WR900207.
Breusch, T.S., 1978. Testing for autocorrelation in dynamic linear models. Aust. Econ. Pap. 17 (31), 334–355. https://doi.org/10.1111/j.1467-8454.1978.tb00635.x.
Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47 (5), 1287–1294.
Brunner, M.I., Melsen, L.A., Newman, A.J., Wood, A.W., Clark, M.P., 2020. Future streamflow regime changes in the united states: assessment using functional classification. Hydrol. Earth Syst. Sci. 24 (8), 3951–3966. https://doi.org/10.5194/hess-24-3951-2020.
Draper, N.R., Guttman, I., 1995. Confidence intervals versus regions. J. R. Stat. Soc. Ser. D (The Statistician) 44 (3), 399–403.
Draper, N.R., Smith, H., 1998. Applied regression analysis. Wiley Series in Probability and Statistics, third ed. John Wiley & Sons, New York, NY, USA.
Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031. https://doi.org/10.1029/91WR02985.
Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression, i. Biometrika 37 (3–4), 409–428. https://doi.org/10.1093/biomet/37.3-4.409.
Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression, ii. Biometrika 38 (1–2), 159–179. https://doi.org/10.1093/biomet/38.1-2.159.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26.
Everitt, B.S., Skrondal, A., 2010. The Cambridge Dictionary of Statistics, fourth ed. Cambridge University Press, New York.
Freer, J., Beven, K., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the glue approach. Water Resour. Res. 32 (7), 2161–2173. https://doi.org/10.1029/95WR03723.
Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., Hochreiter, S., 2020. Data for "rainfall-runoff prediction at multiple timescales with a single long short-term memory network". Zenodo. https://doi.org/10.5281/zenodo.4072701.
Goldfeld, S.M., Quandt, R.E., 1965. Some tests for homoscedasticity. J. Amer. Stat. Assoc. 60 (310), 539–547. https://doi.org/10.1080/01621459.1965.10480811.
Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763. https://doi.org/10.1029/97WR03495.
Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process. 22 (18), 3802–3813. https://doi.org/10.1002/hyp.6989.
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. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. Water Resour. Res. 42 (3) https://doi.org/10.1029/2005WR004368.
Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 2. application. Water Resour. Res. 42 (3) https://doi.org/10.1029/2005WR004376.
Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019. Modular assessment of rainfall–runoff models toolbox (marrmot) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. Geosci. Model Dev. 12 (6), 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019.
Kuczera, G., 1983. Improved parameter inference in catchment models: 1. evaluating parameter uncertainty. Water Resour. Res. 19 (5), 1151–1162. https://doi.org/10.1029/WR019i005p01151.
Kuczera, G., Parent, E., 1998. Monte carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. J. Hydrol. 211 (1), 69–85. https://doi.org/10.1016/S0022-1694(98)00198-X.
Lamontagne, J.R., Barber, C.A., Vogel, R.M., 2020. Improved estimators of model performance efficiency for skewed hydrologic data. Water Resour. Res. 56 (9), 1–25. https://doi.org/10.1029/2020wr027101.
Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116 (12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.
Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part i — a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Newman, A.J., et al., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19 (1), 209–223. https://doi.org/10.5194/hess-19-209-2015.
Pool, S., Vis, M., Seibert, J., 2018. Evaluating model performance: towards a non-parametric variant of the kling-gupta efficiency. Hydrol. Sci. J. 63 (13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002.
Oliveira, D.Y., Vrugt, J.A., 2022. The treatment of uncertainty in diagnostic model evaluation: 1. a probabilistic description of measured streamflow records, Submitted to Water Resources Research.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing (second ed.).
Rakovec, O., Mizukami, N., Kumar, R., Newman, A., Thober, S., Wood, A.W., Clark, M.P., Samaniego, L., 2019. Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous united states. Water Resour. Res. 124 (24), 13991–14007. https://doi.org/10.1029/2019JD030767.
Sadegh, M., Vrugt, J.A., 2013. Bridging the gap between glue and formal statistical approaches: approximate bayesian computation. Hydrol. Earth Syst. Sci. 17 (12), 4831–4850. https://doi.org/10.5194/hess-17-4831-2013.
Scharnagl, B., Iden, S.C., Durner, W., Vereeken, H., Herbst, M., 2015. Inverse modelling of in situ soil water dynamics: accounting for heteroscedastic, autocorrelated, and non-gaussian distributed residuals. Hydrol. Earth Syst. Sci. Discuss. 12, 2155–2199.
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) https://doi.org/10.1029/2009WR008933.
Schoups, G., Vrugt, J.A., Fenicia, F., van de Giesen, N.C., 2010. Corruption of accuracy and efficiency of markov chain monte carlo simulation by inaccurate numerical implementation of conceptual hydrologic models. Water Resour. Res. 46 (10) https://doi.org/10.1029/2009WR008648.
Schwemmle, R., Demand, D., Weiler, M., 2021. Technical note: diagnostic efficiency – specific evaluation of model performance. Hydrol. Earth Syst. Sci. 25, 2187–2198. https://doi.org/10.5194/hess-25-2187-2021.
Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3–4), 591–611. https://doi.org/10.1093/biomet/52.
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. https://doi.org/10.1029/WR016i002p00430.
Spear, R.C., Hornberger, G., 1980. Eutrophication in peel inlet-ii. identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14 (1), 43–49.
Spear, R.C., Cheng, Q., Wu, S.L., 2020. An example of augmenting regional sensitivity analysis using machine learning software. Water Resour. Res. 56 (4), 1–16. https://doi.org/10.1029/2019wr026379.
Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis: 1. ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21 (9), 1421–1432. https://doi.org/10.1029/wr021i009p01421.
Tasker, G.D., 1980. Hydrologic regression with weighted least squares. Water Resour. Res. 16 (6), 1107–1113. https://doi.org/10.1029/wr016i006p01107.
Vogel, R.M., Fennessey, N.M., 1993. L-moment diagrams should replace product-moment diagrams. Water Resour. Res. 29 (6), 1745–1752. https://doi.org/10.1029/93WR00341.
Vrugt, J.A., Beven, K.J., 2018. Embracing equifinality with efficiency: limits of acceptability sampling using the dream(loa) algorithm. J. Hydrol. 559, 954–971. https://doi.org/10.1016/j.jhydrol.2018.02.026.
Vrugt, J.A., Sadegh, M., 2013. Toward diagnostic model calibration and evaluation: approximate bayesian computation. Water Resour. Res. 49 (7), 4335–4345. https://doi.org/10.1002/wrcr.20354.
Vrugt, J.A., Bouten, W., Gupta, H.V., Sorooshian, S., 2002. Toward improved identifiability of hydrologic model parameters: the information content of experimental data. Water Resour. Res. 38 (12) https://doi.org/10.1029/2001WR001118, 48–1–48–13.
Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. Water Resour. Res. 41 (1) https://doi.org/10.1029/2004WR003059.
Vrugt, J.A., Gupta, H.V., Dekker, S.C., Sorooshian, S., Wagener, T., Bouten, W., 2006. Application of stochastic parameter optimization to the sacramento soil moisture accounting model. J. Hydrol. 325 (1), 288–307. https://doi.org/10.1016/j.jhydrol.2005.10.041.
Vrugt, J.A., ter Braak, C., Diks, C., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10 (3), 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273.
Vrugt, J.A., Oliveira, D.Y., Schoups, G., Diks, C.G.H., 2022. On the use of distribution-free likelihood functions: generalized and universal likelihood functions, score rules and multi-criteria ranking. J. Hydrol., submitted.
Westerberg, I.K., Guerrero, J.-L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer, J.E., Xu, C.-Y., 2011. Calibration of hydrological models using flow-duration curves. Hydrol. Earth Syst. Sci. 15 (7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011.
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817–838.
Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the nws distributed hydrologic model. Water Resour. Res. 44 (9) https://doi.org/10.1029/2007WR006716.