# On the use of distribution-adaptive likelihood functions

## Generalized and universal likelihood functions, scoring rules and multi-criteria ranking

Vrugt, Jasper A.; de Oliveira, Debora Y. ; Schoups, Gerrit ; Diks, Cees G.H.

**DOI**
10.1016/j.jhydrol.2022.128542

**Publication date**
2022

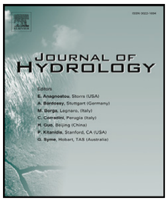
**Document Version**
Final published version

**Published in**
Journal of Hydrology

Contents lists available at ScienceDirect

Journal of Hydrology

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



Research papers

# On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking

Jasper A. Vrugt [a,*], Debora Y. de Oliveira [a], Gerrit Schoups [b], Cees G.H. Diks [c]

[a] *Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA*
[b] *Department of Civil Engineering and Geosciences, Technical University Delft, Delft, The Netherlands*
[c] *Faculty of Economics and Business, Section Quantitative Economics, University of Amsterdam, Amsterdam, The Netherlands*

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Simon Michael Papalexiou, Associate Editor.

*Keywords:*
Bayesian analysis
Distribution-adaptive likelihood
Generalized likelihood
Universal likelihood
Scoring rules
Probabilistic forecasts

ABSTRACT

This paper is concerned with the formulation of an adequate likelihood function in the application of Bayesian epistemology to uncertainty quantification of hydrologic models. We focus our attention on a special class of likelihood functions (hereinafter referred to as distribution-adaptive likelihood functions), which do not require prior assumptions about the expected distribution of the residuals, rather inference takes place over the hypotheses (model parameters) and space of distribution functions. Our goals are threefold. First, we present theory of a revised implementation of the generalized likelihood (GL) function of Schoups and Vrugt (2010) wherein residual standardization precedes the treatment of serial correlation. This so-called $\mathrm{GL}^+$ function, enjoys a solid statistical underpinning and guarantees a more robust joint inference of the autoregressive coefficients and residual properties. Then, as secondary goal, we present a further generalization of the $\mathrm{GL}^+$ function, coined the universal likelihood (UL) function, which extends applicability to highly asymmetrical lepto- and platy-kurtic residual distributions. The UL function builds on the 5-parameter skewed generalized Student's $t$ distribution of Theodossiou (2015) which makes up a large family of continuous probability distributions including (but not limited to) symmetric and skewed forms of the generalized normal, generalized $t$, Laplace, normal, Student's $t$, and Cauchy-Lorentz distributions. As our third and last goal, we present the use of *strictly proper* scoring rules to evaluate, compare and rank likelihood functions. These scoring rules condense the accuracy of a distribution forecast to a single value while retaining attractive statistical properties. The $\mathrm{GL}^+$ and UL functions are illustrated using data of a simple autoregressive scheme and benchmarked against the GL function, Student $t$ likelihood (SL) of Scharnagl et al. (2015) and normal likelihood (NL) for a conceptual hydrologic model using measured streamflow data. Our results show that, (i) the $\mathrm{GL}^+$ function is superior to the GL function, (ii) the active set of nuisance variables exerts a large control on the performance of the $\mathrm{GL}^+$, SL and UL functions, (iii) the treatment of autocorrelation deteriorates the scoring rules and performance metrics of the forecast distribution, (iv) a leptokurtic distribution is favored for discharge residuals, (v) scoring rules are indispensable in our search for the true forecast distribution, and (vi) the use of multiple *strictly proper* scoring rules turns the selection of an adequate likelihood function into a multi-criteria problem.

## 1. Introduction and scope

In the past decades much progress has been made in uncertainty quantification of the parameters, state variables and simulated output of dynamic system models. In all this work, Bayes theorem has emerged as a corner stone of modern probability theory (hypothesis testing) and as working paradigm for the subjectivist approach to epistemology, statistics, and inductive logic. Bayesian inference allows for an exact description of parameter uncertainty (and other sources of uncertainty) by treating the parameters as probabilistic variables with joint posterior probability density function (PDF). This so-called posterior parameter distribution, $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, is the consequence of a prior distribution, $p(\boldsymbol{\theta})$, which encodes all "subjective" knowledge about the parameters, $\boldsymbol{\theta} = [\,\theta_1\ \theta_2\ \dots\ \theta_d\,]^\top$, before collection of the data, $\widetilde{\mathbf{y}}$, and a likelihood, $L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, which was designated as mathematical quantity by Fisher (1934) to quantify our degree of belief (confidence) in the parameter values, $\boldsymbol{\theta}$, in light of the observed data, $\widetilde{\mathbf{y}}$. Bayes (1763) theorem expresses the mathematical relationship between the prior, $p(\boldsymbol{\theta})$, likelihood, $L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, and posterior (= updated) beliefs, $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, of the

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

https://doi.org/10.1016/j.jhydrol.2022.128542
Received 7 March 2022; Received in revised form 16 August 2022; Accepted 24 September 2022
Available online 9 November 2022

parameters

$$p(\boldsymbol{\theta}|\widetilde{\mathbf{y}}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})} \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}), \tag{1}$$

where the denominator, $p(\widetilde{\mathbf{y}}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, acts as a normalizing constant so that the posterior parameter distribution integrates to unity over the prior parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$. This so-called model evidence or marginal likelihood plays a crucial role in hypothesis testing when selecting the best model among a collection of competing models (Volpi et al., 2017) but can be ignored during parameter inference as the unnormalized posterior density, $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, suffices to estimate $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$. The $100(1-\alpha)\%$ credible region, $\mathcal{C}_\alpha \subseteq \boldsymbol{\Theta}$, of the vector-valued, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, will make up a $(1-\alpha)$ proportion of the probability mass of $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$, at a significance level, $\alpha \in (0,1)$. Credible intervals, $\mathcal{C}_\alpha = (a,b)$, of each parameter, $\theta_j;\ j = (1,2,\dots,d)$, may be construed separately

$$\int_{-\infty}^{a} p(\theta_1,\theta_2,\dots,\theta_j,\dots,\theta_d|\widetilde{\mathbf{y}})\mathrm{d}\boldsymbol{\theta}_{-j} = \int_{b}^{\infty} p(\theta_1,\theta_2,\dots,\theta_j,\dots,\theta_d|\widetilde{\mathbf{y}})\mathrm{d}\boldsymbol{\theta}_{-j} = \alpha/2, \tag{2}$$

where marginalization takes places over $\boldsymbol{\theta}_{-j}$, the parameter space without $\theta_j$. Bayesian credible regions and/or intervals are analogous to the confidence regions/intervals of generalized least squares (GLS) used in frequentist statistics (Lee, 2012), although subtle but important differences exist in their computation and interpretation (Lu et al., 2012). For a multi-modal posterior parameter distribution, confidence/credible intervals are rather meaningless and we should resort instead to a highest posterior density (HPD) region. A HPD region, $\mathcal{H}_\alpha$, of $\alpha$ significance level is a $(1-\alpha)$ region for which holds that $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}}) \geq p(\boldsymbol{\theta}'|\widetilde{\mathbf{y}})$ for all $\boldsymbol{\theta} \in \mathcal{H}_\alpha$ and $\boldsymbol{\theta}' \in \boldsymbol{\Theta}\backslash\mathcal{H}_\alpha$, where $A\backslash B = \{x \in A \text{ and } x \notin B\}$. After we have processed the data, $\widetilde{\mathbf{y}}$, we can evaluate the posterior predictive PDF of the simulated quantity, $y(\boldsymbol{\theta}) \in \mathbb{R}$

$$p(y|\widetilde{\mathbf{y}}) = \int_{\boldsymbol{\Theta}} p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})\mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where $p(y|\boldsymbol{\theta})$ is the density function of the distribution of $y$ given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This predictive distribution reflects both the aleatory uncertainty in future observations and the posterior uncertainty in the parameters and is directly related to its CDF, $F(y|\widetilde{\mathbf{y}})$

$$F(y|\widetilde{\mathbf{y}}) = \int_{\boldsymbol{\Theta}} F_{\mathrm{c}}(y|\boldsymbol{\theta})\mathrm{d}p(\boldsymbol{\theta}|\widetilde{\mathbf{y}}), \tag{4}$$

where $F_{\mathrm{c}}(y|\boldsymbol{\theta})$ is the *conditional* predictive distribution when $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ equal the true parameter values (Greenberg, 2013; Gelman et al., 2014). Unfortunately, the above integrals do not often admit a solution in closed form, and, thus, must be approximated using some form of Monte Carlo simulation (Gelfand and Smith, 1990; Gilks et al., 1996).

The likelihood function exerts a large control on the posterior parameter and predictive distributions, $p(\boldsymbol{\theta}|\widetilde{\mathbf{y}})$ and $p(y|\widetilde{\mathbf{y}})$, respectively, and has been the subject of much debate in the hydrologic literature (Beven and Binley, 1992; Freer et al., 1996; Yang et al., 2007; Smith et al., 2008; Smith and Marshall, 2008; Reichert and Mieleitner, 2009; Vrugt et al., 2009; Schoups and Vrugt, 2010; Smith et al., 2010; Evin et al., 2013; Scharnagl et al., 2015; Smith et al., 2015; Li et al., 2016; McInerney et al., 2019; Ammann et al., 2019). These contributions may be grouped in two main camps with fundamentally different and opposing viewpoints on how to characterize uncertainty in the presence of epistemic errors and residual non-stationarity. The first camp believes in the power of statistical theory and treats uncertainty as aleatory in nature. Proponents of this camp insist that likelihood functions should have a solid theoretical underpinning, but despite many efforts to understand and describe epistemic (systematic) uncertainty, are often silent on how one should characterize and treat model structural errors and non-stationary residual behavior. The residuals display certain patterns, and it is in the features of these patterns that we may distill a suitable likelihood function with roots in probability theory. This camp of idealists is most likely to include theoreticians and hydrologists with a strong STEM education. Contributions include the use of Eq. (6) and variations thereof Sorooshian and Dracup (1980), Tasker (1980), Kuczera (1982), Stedinger and Tasker (1985), Bates and Campbell (2001), Smith et al. (2010), Evin et al. (2013), McInerney et al. (2019), the application of advanced distribution-adaptive likelihood functions (Reichert and Mieleitner, 2009; Schoups and Vrugt, 2010; Scharnagl et al., 2015; Ammann et al., 2019) and implementation of set-theoretic likelihood functions within the context of approximate Bayesian computation (Sadegh and Vrugt, 2013). The second camp recognizes the (current) limitations of formal statistical theory in describing systematic uncertainty and takes on an alternative philosophy and approach to model evaluation. This includes the use of so-called pseudo-likelihood functions within the GLUE methodology (Beven and Binley, 1992; Freer et al., 1996; Beven and Freer, 2001; Smith et al., 2008), set-theoretic likelihood functions within the context of Limits of Acceptability (Beven, 2006; Vrugt and Beven, 2018) and voting-point likelihoods for rating curve estimation in the presence of aleatory and epistemic uncertainty (McMillan and Westerberg, 2015; Hollaway et al., 2018). This camp of pragmatists will include practicing hydrologists who seek ways to formalize their field expertise and knowledge of watershed characteristics and behavior into expectations about model performance.

In this paper, we take a formal probabilistic viewpoint and focus our attention on so-called distribution-adaptive likelihood functions. This special group of likelihood functions does not make prior assumptions about the expected distribution of the residuals. Rather, inference takes place over the model parameters, $\boldsymbol{\theta}$, and space of distribution functions defined by the shape parameters of the likelihood function. We label this group distribution-adaptive likelihood functions after distribution-free maximum likelihood (Cosslett, 1983) and partially-adaptive (Zeckhauser and Thompson, 1970; Hansen et al., 2006) estimation. The goals of this paper are threefold. First, we present theory of a revised implementation of the generalized likelihood (GL) function of Schoups and Vrugt (2010). This new formulation, referred to as the GL$^+$ function, rectifies a critical deficiency of the GL function pointed out by Evin et al. (2013) and guarantees a (much) more robust joint inference of the autoregressive coefficients and distribution of the (partial) residuals. Then, as secondary goal, we present a further generalization of the GL$^+$ function, coined the universal likelihood (UL) function. The UL function builds on the skewed generalized Student's $t$-distribution of Theodossiou (2015) and extends applicability to a much larger family of continuous probability distributions among which highly asymmetrical lepto- and platy-kurtic residual distributions. This includes (but is not limited to) symmetric and skewed forms of the generalized normal, generalized $t$ (McDonald and Newey, 1988), Laplace (Laplace, 1774), normal, Student's $t$ (Student, 1908), Cauchy-Lorentz (Poisson, 1824) and uniform distribution. Finally, as our third and last goal, we present the use of *strictly proper* scoring rules to evaluate, compare and rank the probabilistic forecasts of the likelihood functions. Scoring rules are a powerful alternative to more intuitive metrics such as the reliability, coefficient of variation and coverage and condense the accuracy of a distribution forecast to a single penalty oriented value while retaining attractive statistical properties (Alexander et al., 2022). The GL$^+$ and UL functions are tested using synthetic data of a simple autoregressive scheme and benchmarked against the GL function, Student $t$ likelihood (SL) of Scharnagl et al. (2015) and normal likelihood for a conceptual hydrologic model using measured streamflow data.

The remainder of this paper is organized as follows. Section 2 provides a problem statement and reviews the definition of the default (normal) likelihood function. Section 3 presents theory and a derivation of the GL$^+$ function. In this section we also revisit the skewed Student $t$ distribution of Scharnagl et al. (2015) which serves as central ingredient of the so-called Student likelihood (SL) function. This is followed in Section 4 by a detailed description of the UL function. Then, Section 5 presents the scoring rules used to evaluate, compare and rank the predictive distributions derived from the different likelihood functions. This is followed by the application of the GL$^+$ and UL functions to two different case studies involving a simple autoregressive scheme and conceptual watershed model. We conclude this paper in Section 6 with a summary of our main findings. To be complete in our reasoning and arguments and convey the subtleties involved in the application, evaluation and ranking of distribution-adaptive likelihood functions, this paper is relatively heavy on statistics. This is a necessary means to describing accurately the distribution of the residuals and prerequisite to diagnosing and detecting model structural errors in an effort to enhance our knowledge and understanding of the rainfall–discharge relationship.

## 2. Problem statement and background

Let us consider a deterministic system, $\mathfrak{S}$, whose future state, behavior and regularities may be described by one or more rules or principles of change, which may be spatially coupled and explain its temporal evolution. As most real-world systems have an intractable complexity, our understanding of the dynamics of the system is limited to one or more variables whose values have been measured in situ or sensed remotely over a period of time. For simplicity, we consider a single variable, $\chi$, whose behavior is measured and collected in a $n \times 1$ vector, $\widetilde{\mathbf{y}} = [\, \widetilde{y}_1 \; \widetilde{y}_2 \; \dots \; \widetilde{y}_n \,]^\top$, where the tilde operator denotes measured quantities.

If, the real-world system, $\mathfrak{S}$, is considered deterministic then we must be able to describe its behavior with explanatory laws, for example, those most fundamental and celebrated in physics. Let us suppose that the dynamic model, $\mathbf{y} \leftarrow \mathcal{M}(\boldsymbol{\theta}, \mathbf{X}) \; : \; \mathbb{R}^d \; \rightarrow \; \mathbb{R}^n$, simulates the temporal behavior, $\mathbf{y} = [\, y_1 \; y_2 \; \dots \; y_n \,]^\top$, of the variable $\chi$ of interest using as input arguments a $d$-vector of unknown invariant system properties, $\boldsymbol{\theta} = [\, \theta_1 \; \theta_2 \; \dots \; \theta_d \,]^\top$, with $\boldsymbol{\theta} \in \mathbb{R}^d$, and array, $\mathbf{X}$, of constants and input variables required under the supposition or hypothesis that they govern, by causality, the variable $\chi$ using the evolution rules of nature expressed in mathematical form. These physical laws are the result of thought experiments, mathematical analysis and derivation, and laboratory and/or field experimentation under idealized external (boundary) conditions. A deterministic or metaphysical model scales up the physical laws, a mixture of algebraic, ordinary and partial differential equations, to the domain of the real-world system of interest to warrant spatiotemporal simulation of the governing variables of interest at the desired space–time resolution. In some cases the evolution rule can be simplified to a mathematical expression (Van Geert, 1994) but for most real-world systems the dynamics are simply too complex to warrant an exact characterization by one or more mathematical functions. The array, $\mathbf{X}$, characterizes the system's initial state, invariant properties and spatiotemporal control inputs (forcing/explanatory variables). As our main focus is on the model parameters, $\boldsymbol{\theta}$, we suppress the symbol $\mathbf{X}$ in our subsequent notation and write instead, $\mathbf{y} = \mathcal{M}(\boldsymbol{\theta})$, 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 measured response, $\widetilde{\mathbf{y}} = [\, \widetilde{y}_1 \; \widetilde{y}_2 \; \dots \; \widetilde{y}_n \,]^\top$ of variable $\chi$. As the data generating system, $\mathfrak{S}$, is considered deterministic and described by laws of nature there is no randomness in the evolution of future behavior. Any uncertainty in $\chi$ is due to measurement error only, an inherent byproduct of the measurement process. If we make the common assumption that the measurement errors of $\chi$ are random and additive

$$\widetilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\epsilon}, \tag{5}$$

then the $n$-vector, $\mathbf{y} = [\, y_1 \; y_2 \; \dots \; y_n \,]^\top$, equals the unobserved "true" response of the deterministic system, $\mathfrak{S}$, which the model, $\mathcal{M}(\cdot)$, is supposed to mimic. This paper is concerned with a formal probabilistic description of the residuals, $\mathbf{e}(\boldsymbol{\theta}) = \widetilde{\mathbf{y}} - \mathcal{M}(\boldsymbol{\theta})$, so as to help guide the inference to the unobserved "true" response, $\mathbf{y}$.

To quantify the level of confidence (aka likelihood) in the simulated outcome, $y_t(\boldsymbol{\theta})$, given observation, $\widetilde{y}_t$, we must hypothesize a distribution, $f_{\widetilde{y}_t}$, of the measurement, $\widetilde{y}_t$. As the data generating process, $\mathfrak{S}$, is assumed deterministic, there is no randomness in the system state and behavior at any moment in time. Then, Eq. (5) implies that $\widetilde{y}_t$ should follow the distribution of the measurement error, $\epsilon_t$. If we make the common and convenient assumption that $e_t(\boldsymbol{\theta})$ is zero-mean normally distributed with variance, $\sigma_e^2$, and, thus, $e_t(\boldsymbol{\theta}) \; \sim \; \mathcal{N}(0, \sigma_e^2)$, then the distribution of the measurement, $f_{\widetilde{y}_t} \; = \; f_{\mathcal{N}}(\cdot|\widetilde{y}_t, \sigma_e^2)$, where $f_{\mathcal{N}}(a|b,c)$ signifies the PDF of the normal distribution, $\mathcal{N}(b,c)$, with mean, $b$, and variance, $c$, evaluated at the simulated outcome, $a$. The likelihood of the parameters now equals the density of the normal PDF at the simulated outcome, to yield, $L(\boldsymbol{\theta}|\widetilde{y}_t, \sigma_e^2) = f_{\mathcal{N}}\big(y_t(\boldsymbol{\theta})|\widetilde{y}_t, \sigma_e^2\big)$. We can generalize the computation of the likelihood to a $n$-vector of simulated output. This leads to the well-known GLS likelihood function, $L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\Sigma}_e)$, and equals

$$L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\Sigma}_e) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_e|}} \exp\Big( -\frac{1}{2} \mathbf{e}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}_e^{-1} \mathbf{e}(\boldsymbol{\theta}) \Big) \tag{6}$$

where $\boldsymbol{\Sigma}_\epsilon$ is the $n \times n$ measurement error covariance matrix, $|\cdot|$ is the determinant operator, $n$ denotes the length of the training data record and the multiplicative term in front of the exponential function is the familiar normalizing constant of a normal distribution. The maximum, $\widehat{\boldsymbol{\theta}}$, of $L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\Sigma}_e)$ is also known as the maximum likelihood solution.

The expectation that the measurement errors of the data generating process will provide an exhaustive description of the residuals is flawed. As complex systems do not admit a perfect characterization, the residuals, $\mathbf{e}(\boldsymbol{\theta}) \; = \; \widetilde{\mathbf{y}} - \mathbf{y}(\boldsymbol{\theta})$, will almost always be larger than the measurement errors, $\boldsymbol{\epsilon}$, of the measured response, $\widetilde{\mathbf{y}}$. The residuals are simply expected to absorb the consequences of model misspecification and inadequate characterization of the initial states, system properties and controlling variables and behave analogously to the measurement errors of $\widetilde{\mathbf{y}}$. This is equivalent to a residual covariance matrix, $\boldsymbol{\Sigma}_e$, which is a multiple of the measurement error covariance matrix, $\boldsymbol{\Sigma}_\epsilon$. In the case of a scalar covariance matrix, we can write, $\boldsymbol{\Sigma}_e \; = \; c\sigma_\epsilon^2 \mathbf{V}$, where $\sigma_e^2 \; = \; c\sigma_\epsilon^2$ signifies the (unknown) variance of the residuals. Thus, in practice, we must submit instead to Eq. (6) the $n \times n$ covariance matrix of the residuals, $\boldsymbol{\Sigma}_e$, so that the standardized and decorrelated residuals satisfy the assumptions of a zero-mean and constant variance of the Gauss–Markov theorem. We can slip into usage of the terminology measurement error, but this characterization is imprecise in the present context (e.g. see Page 20 of Beven, 2006).

The multivariate form of the GLS likelihood in Eq. (6) permits treatment of residual autocorrelation through the off-diagonal terms of the residual covariance matrix, $\boldsymbol{\Sigma}_e$. For example, for a first-order autoregressive process, $e_t(\boldsymbol{\theta}) \; = \; \phi_1 e_{t-1}(\boldsymbol{\theta}) + \varepsilon_t$, with zero-mean normally distributed partial residuals, $\varepsilon_t \; \sim \; \mathcal{N}(0, \sigma_\varepsilon^2)$, with variance, $\sigma_\varepsilon^2$, the residual covariance matrix, $\boldsymbol{\Sigma}_e$, may be factorized to a product, $\boldsymbol{\Sigma}_e = \big(\mathbf{L}(\phi_1)\mathbf{U}(\phi_1, \sigma_\varepsilon^2)\big)^{-1}$, of a lower and upper triangular matrix, $\mathbf{L}(\phi_1)$ and $\mathbf{U}(\phi_1, \sigma_\varepsilon^2)$, respectively, which depend only on the autoregressive coefficient, $\phi_1$, and/or $\sigma_\epsilon^2$ and the determinant, $|\boldsymbol{\Sigma}_e| = \sigma_\epsilon^{2n}/(1-\phi_1^2)$. Unfortunately, not all surmised univariate residual distributions support a closed-form multivariate expression for the joint likelihood. This is particularly true for the nontraditional residual distributions discussed herein. Then, we must treat serial correlation prior to computation of the marginal and joint likelihoods.

## 3. Generalized likelihood function

The GLS likelihood function has found widespread application and use, yet, involves the rather questionable assumption that the residuals are normally distributed. Power transformations such as the (Box and Cox, 1964) transform may help stabilize the variance of the data and model simulations and make their respective marginal distributions more normal distribution-like, yet the power and/or shift parameters of this transformation are not unambiguously defined (McInerney et al., 2017) and the transformation does not necessarily protect against heavy-tailed residuals (Bates and Campbell, 2001; Yang et al., 2007). Distribution-adaptive likelihood functions such as the GL function of Schoups and Vrugt (2010) allow us to relax the assumption of normality and describe any arbitrary residual distribution with (non)constant variance and/or different degrees of skew and kurtosis. Unlike the Box–Cox transformation which simultaneously affects several moments of the residual distribution (Ammann et al., 2019), the GL function separately treats residual heteroscedasticity, skewness and kurtosis with parameters that have a direct relation to residual statistics and can be deduced from diagnostic plots (Schoups and Vrugt, 2010). This not only lets the measured data, $\widetilde{\mathbf{y}}$, speak for itself, but also leads to a more robust characterization of model parameter and predictive uncertainty. A skewed residual distribution will, for instance, improve the probabilistic description of truncated variables, say, near-zero flows in a semiarid watershed (Schoups and Vrugt, 2010).

In this section, we present a revised formulation of the GL function of Schoups and Vrugt (2010) which enacts treatment of serial correlation on homogenized rather than non-homogenized (heteroscedastic) residuals. This revised formulation simplifies the joint inference of the parameters, $\boldsymbol{\theta}$, and probabilistic properties of the residuals. The theoretical derivation below serves as precursor to the UL function which is presented in a subsequent section.

### 3.1. The standardized skew exponential power (SEP) density function

The GL function of Schoups and Vrugt (2010) builds on the standardized skew exponential power (SEP) density function which combines the distribution of Subbotin (1923), also known as the exponential power (EP) or generalized normal distribution (Box and Tiao, 1992)

$$f_{\mathrm{EP}}(a|0,1,\beta) = \omega_\beta \exp\left(-c_\beta |a|^{2/(1+\beta)}\right), \tag{7}$$

and the template density, $f_{\mathrm{skew}}(a|\xi)$, for skewed distributions of Fernandez and Steel (1998)

$$f_{\mathrm{skew}}(a|\xi) = \frac{2}{\xi + \xi^{-1}} f\left(a\xi^{-\,\mathrm{sign}(a)}\right), \tag{8}$$

to yield (see Appendix A)

$$f_{\mathrm{SEP}}(a|0,1,\beta,\xi) = \frac{2\sigma_\xi \omega_\beta}{\xi + \xi^{-1}} \exp\left(-c_\beta \left| \frac{\mu_\xi + \sigma_\xi a}{\xi^{\mathrm{sign}(\mu_\xi + \sigma_\xi a)}} \right|^{2/(1+\beta)}\right), \tag{9}$$

where $|\cdot|$ denotes the absolute value or modulus operator, $\mathrm{sign}(x) = |x|/x$, is the signum function and the scalars $c_\beta$, $\omega_\beta$, $\mu_\xi$, and $\sigma_\xi$, are a function of the kurtosis, $\beta \in (-1, 1]$, and skewness, $\xi \in \mathbb{R}_+$, using Eqs. (A.2a), (A.2b), (A.7a) and (A.7b), respectively. Appendix A completes the derivation of the SEP density in Schoups and Vrugt (2010) with a detailed mathematical treatment of the mean, $\mu_\xi$, and variance, $\sigma_\xi^2$, of the SEP distribution.

To provide insights into the functional shape of the standardized SEP density of Eq. (9), please consider Fig. 1 which presents a graph of $f_{\mathrm{SEP}}(a|0,1,\beta,\xi)$ for $a \in [-3, 3]$ using different values of the (a) kurtosis, $\beta$, and (b) skewness, $\xi$.

As is evident from the two graphs, the density of the standardized SEP distribution at point $a$ is controlled by the values of $\beta$ and $\xi$. The density is symmetric for $\xi = 1$, positively skewed for $\xi > 1$ and negatively skewed for $\xi < 1$. For a symmetric density, that is $\xi = 1$, a value of $\beta = -1$ results in a uniform distribution, $\beta = 0$ produces a normal distribution, and $\beta = 1$ equals a double-exponential or Laplace distribution. Thus, a value of $\beta \in (0, 1]$ produces a SEP distribution with (much) heavier tails than the normal density. Such leptokurtic distributions help protect the model parameter estimates from outliers.

### 3.2. Treatment of residual autocorrelation

The standardized SEP density of Eq. (9) enables a fluent description of the distribution of the residuals. Before we can proceed, however, with the derivation of the SEP likelihood function, we are in need of an adequate description of the temporal structure of the raw residuals, $\mathbf{e}(\boldsymbol{\theta})$. Schoups and Vrugt (2010) expressed serial correlation as an AR($k$) process of the raw residuals, $\mathbf{e}(\boldsymbol{\theta})$, as follows

$$e_t(\boldsymbol{\theta}) = \sum_{j=1}^{k} \phi_j e_{t-j}(\boldsymbol{\theta}) + \sigma_{e_t} \varepsilon_t, \tag{10}$$

where $\boldsymbol{\Phi}_k = [\,\phi_1 \; \phi_2 \; \ldots \; \phi_k\,]^\top$ is a $k$-vector of autoregressive coefficients, $\sigma_{e_t}$, signifies the standard deviation of the $t^{\text{th}}$ raw residual, $\varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta})$, denotes the AR($k$)-decorrelated residual and, $\boldsymbol{\delta} = [\,\beta \;\; \xi \;\; \boldsymbol{\Phi}_k\,]^\top$, is a vector of nuisance variables.[1] The entries of the $n \times 1$ vector, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\delta}) = [\,\varepsilon_1(\boldsymbol{\theta}, \boldsymbol{\delta}) \;\; \varepsilon_2(\boldsymbol{\theta}, \boldsymbol{\delta}) \;\; \ldots \;\; \varepsilon_n(\boldsymbol{\theta}, \boldsymbol{\delta})\,]^\top$, are also referred to as innovations or partial residuals.

The autoregressive operator in Eq. (10) operates on the non-standardized raw residuals, $\mathbf{e}(\boldsymbol{\theta})$. This implementation is not recommended as the autoregressive operator of Eq. (10) expects use of homogenized residuals. Evin et al. (2013) convincingly demonstrated that residual standardization should precede the treatment of serial correlation. This guarantees a more robust inference of the model parameters and/or nuisance variables, and, consequently, may lead to sharper prediction intervals. Therefore, and as in Steinschneider et al. (2015) and Hernández-López and Francés (2017), we admit standardized raw residuals, $\underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta})$, with unknown mean and variance, $\sigma_{\underline{e}}^2$, instead

$$\underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{e_t(\boldsymbol{\theta})}{\sigma_{\epsilon_t}} \quad \forall\, t = (1, 2, \ldots, n), \tag{11}$$

and conveniently limit our mathematical description of their structural dependence to an AR(2)-process

$$\underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) = \phi_1 \underline{e}_{t-1}(\boldsymbol{\theta}, \boldsymbol{\delta}) + \phi_2 \underline{e}_{t-2}(\boldsymbol{\theta}, \boldsymbol{\delta}) + \varepsilon_t, \tag{12}$$

where the partial residuals, $\varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta}) \sim \mathcal{SEP}(0, \sigma_\varepsilon^2, \beta, \xi)$, are assumed to be independent and zero-mean SEP distributed with variance, $\sigma_\varepsilon^2$, skewness, $\beta$, and kurtosis, $\xi$, for all $t = (3, 4, \ldots, n)$ and the single line underneath the raw residuals articulates their standardized dimensionless counterparts, $\underline{\mathbf{e}}(\boldsymbol{\theta}, \boldsymbol{\delta}) = [\,\underline{e}_1(\boldsymbol{\theta}, \boldsymbol{\delta}) \;\; \underline{e}_2(\boldsymbol{\theta}, \boldsymbol{\delta}) \;\; \ldots \;\; \underline{e}_n(\boldsymbol{\theta}, \boldsymbol{\delta})\,]^\top$. Note that the standardized raw residuals support the use of more elaborate AR models, for example, with seasonally-varying coefficients or AR-coefficients that depend on exogenous variables (Ammann et al., 2019). This is not uncommon in statistical hydrology, and could be advised if model performance depends on time (season) and/or the governing processes.

The AR(2)-process of Eq. (12) should typically suffice in removing serial dependencies of the standardized raw residuals, otherwise, one could always implement a higher-order AR($k$)-process with $k > 2$. Five remarks are in order. First, the recursion in Eq. (12) demands specification of so-called initial conditions, $\underline{e}_{-1}(\boldsymbol{\theta}, \boldsymbol{\delta})$ and/or $\underline{e}_0(\boldsymbol{\theta}, \boldsymbol{\delta})$, for computation of the first two partial residuals, $\varepsilon_1(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\varepsilon_2(\boldsymbol{\theta}, \boldsymbol{\delta})$. These initial conditions not only determine the values of $\varepsilon_1(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\varepsilon_2(\boldsymbol{\theta}, \boldsymbol{\delta})$ but also their respective marginal distributions. A second and related comment, the moments of the SEP innovations, $\varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta})$, are not invariant under the transformation of Eq. (12). Thus, the standardized raw residuals, $\underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta})$, do not inherit the variance, $\sigma_\varepsilon^2$, skewness, $\beta$, and kurtosis, $\xi$, of the partial residuals. A change of variance, thus, $\sigma_{\underline{e}}^2 \neq \sigma_\varepsilon^2$, is even expected with normally distributed innovations, but the variant skew and kurtosis are the byproduct of the use of non-Gaussian innovations. Third, Eq. (11) does not guarantee a zero mean and unit variance of the standardized raw residuals. In fact, the

[1] A nuisance variable is fundamental to the probabilistic model of concern which is not of immediate interest but must be accounted for in the inference

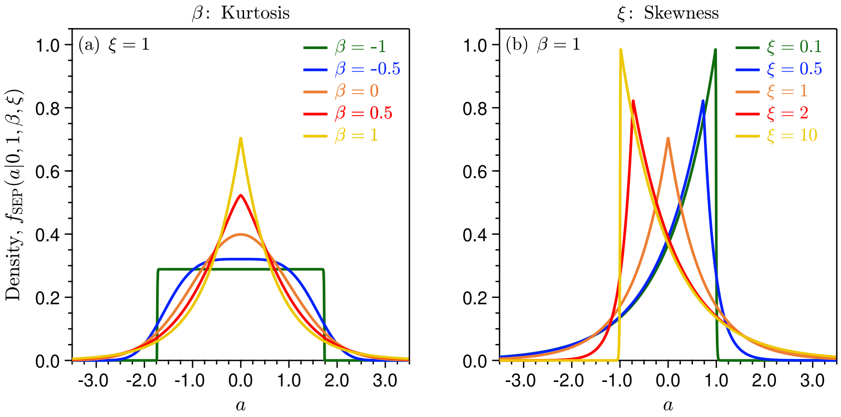


**Fig. 1.** Density of the standardized skew exponential power (SEP) distribution, $f_{\text{SEP}}(a|0,1,\beta,\xi)$, of Eq. (9) for different values of (a) the kurtosis, $\beta$, using $\xi = 1$ and (b) skewness, $\xi$, using $\beta = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variance of the standardized raw residuals will almost always be larger than one as the magnitude of the $e_t(\theta,\delta)$'s will exceed, on average, the measurement error standard deviations, $\sigma_{\epsilon_t}$, of the training data, $\widetilde{y}_t$, for $t = (1,2,\dots,n)$. We can inflate the $\sigma_{\epsilon_t}$'s so as to absorb all other sources of systematic and aleatory uncertainty and impose a unit variance of the standardized raw residuals, $\sigma^2_{\underline{e}} = 1$. In this case, we should replace the $\sigma_{\epsilon_t}$'s with $\sigma_{e_t}$, the standard deviation of the $t^{\text{th}}$ residual. Fourth, it is quite common to replace the denominator of Eq. (11) with estimates, $s_{\epsilon_t}$, of the measurement error standard deviation, $\sigma_{\epsilon_t}$. In doing so, the $\underline{e}_t(\theta,\delta)$'s in Eq. (11) should be referred to as *studentized* residuals.[2] We will use the terminology *studentized residual* throughout the remainder of this paper as this characterizes better the application of Eq. (11). Lastly, the use of $s_{\epsilon_t}$ rather than $\sigma_{\epsilon_t}$ in the denominator of Eq. (11) associates the nuisance variables, $\delta$, to the studentized raw residuals, in a manner that will become explicit in Section 3.4. Therefore, we write, $\underline{\mathbf{e}}(\theta,\delta) = [\,\underline{e}_1(\theta,\delta)\ \underline{e}_2(\theta,\delta)\ \dots\ \underline{e}_n(\theta,\delta)\,]^\top$, for the studentized raw residuals.

Next, we must establish a relationship between the variance, $\sigma^2_\varepsilon$, of the innovations, $\varepsilon_t(\theta,\delta)$, and the unconditional (or marginal) variance, $\sigma^2_{\underline{e}}$, of the studentized raw residuals. The derivation of this relationship can be found in statistical text books and relies on variance decomposition of Eq. (12) to yield

$$\sigma^2_\varepsilon = \frac{(1+\phi_2)(1-\phi_1-\phi_2)(1+\phi_1-\phi_2)}{(1-\phi_2)}\sigma^2_{\underline{e}}. \tag{13}$$

The quotient on the right-hand side is on the unit interval and, thus, we yield that $\sigma^2_{\underline{e}} > \sigma^2_\varepsilon$, unless $\phi_1 = 0$ and $\phi_2 = 0$, then, $\sigma^2_\varepsilon = \sigma^2_{\underline{e}}$. If $\phi_2 = 0$, the above expression reduces to $\sigma^2_\varepsilon = (1-\phi_1^2)\sigma^2_{\underline{e}}$, the familiar expression for the relationship between the variance of an AR(1) process, $\sigma^2_{\underline{e}}$, and the variance, $\sigma^2_\varepsilon$, of the partial residuals.

### 3.3. Derivation of the SEP likelihood function

If we consider the values of the $s$ nuisance variables, $\delta = [\,\beta\ \xi\ \boldsymbol{\Phi}_2\,]^\top$, to be known a-priori then the joint likelihood, $L(\theta|\widetilde{\mathbf{y}},\delta)$, of the $d$ model parameters, $\theta = [\,\theta_1\ \theta_2\ \dots\ \theta_d\,]^\top$

$$L(\theta|\widetilde{\mathbf{y}},\delta) \equiv p(\widetilde{\mathbf{y}},\delta|\theta) = p(\mathbf{e},\delta|\theta), \tag{14}$$

equals the joint PDF of the training data, $\widetilde{\mathbf{y}}$, $\delta$ and $\theta$. To compute the likelihood of the $n$-vector of raw residuals, $\mathbf{e}(\theta)$, we must make assumptions about the so-called initial conditions of the AR(2) model in Eq. (12). The initial conditions, $e_{-1}(\theta) = \widetilde{y}_{-1} - y_{-1}(\theta)$ and $e_0(\theta) = \widetilde{y}_0 - y_0(\theta)$, not only determine, after studentization with Eq. (11), the values of the first two partial residuals, $\varepsilon_1(\theta,\delta)$ and $\varepsilon_2(\theta,\delta)$, but also control the next successive entries of $\varepsilon(\theta,\delta)$ via the recursion in Eq. (12). If the measured data, $\widetilde{y}_{-1}$ and $\widetilde{y}_0$, and model output, $y_{-1}(\theta)$ and $y_0(\theta)$, are known at $t = -1$ and $t = 0$ immediately preceding the first entry of the training data record, then the so-called initial conditions equal, $e_{-1} = \widetilde{y}_{-1} - y_{-1}(\theta)$ and $e_0 = \widetilde{y}_0 - y_0(\theta)$. Otherwise, the user can fix the raw residuals at some default value, say, $e_{-1}(\theta) = 0$ and $e_0(\theta) = 0$. In this latter case, we do not adapt our notation and carry forward the dependence of the initial condition on the assumed parameter values.

The joint distribution of the raw residuals, $p(\mathbf{e},\delta|\theta)$, satisfies the following identity

$$p(\mathbf{e}|\theta) = p(\underline{\mathbf{e}}|\theta,\delta)\left|\frac{\partial\underline{\mathbf{e}}(\theta,\delta)}{\partial\mathbf{e}(\theta)}\right|, \tag{15}$$

where $p(\underline{\mathbf{e}}|\theta,\delta)$ is the joint distribution of the studentized raw residuals and, $\partial\underline{\mathbf{e}}(\theta,\delta)/\partial\mathbf{e}(\theta)$, signifies the Jacobian of the transformation from the studentized raw residuals, $\underline{\mathbf{e}}(\theta,\delta)$, to the raw residuals, $\mathbf{e}(\theta)$. The joint distribution of the studentized raw residuals, $p(\underline{\mathbf{e}}|\theta,\delta)$, equals the product of the marginal densities, $p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)$ and $p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)$, of $\underline{e}_1(\theta,\delta)$ and $\underline{e}_2(\theta,\delta)$, respectively, and the conditional densities, $p(\underline{e}_t|\underline{e}_{t-1},\underline{e}_{t-2};\theta,\delta)$, of the remaining $n-2$ entries, $\{\underline{e}_3(\theta,\delta),\dots,\underline{e}_n(\theta,\delta)\}$, of $\underline{\mathbf{e}}(\theta,\delta)$, to yield

$$\begin{aligned}
p(\underline{\mathbf{e}}|\theta,\delta) &= p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)p(\underline{e}_3,\dots,\underline{e}_n|\underline{e}_0,\underline{e}_{-1};\theta,\delta)\\
&= p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)\prod_{t=3}^{n}p(\underline{e}_t|\underline{e}_{t-1},\dots,\underline{e}_0,\underline{e}_{-1};\theta,\delta)\\
&= p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)\prod_{t=3}^{n}p(\underline{e}_t|\underline{e}_{t-1},\underline{e}_{t-2};\theta,\delta).
\end{aligned} \tag{16}$$

If we make the convenient assumption that the partial residuals, $\varepsilon(\theta,\delta)$, follow a normal distribution with zero-mean and variance, $\sigma^2_\varepsilon$, then the marginal distributions of $\underline{e}_1(\theta,\delta)$ and $\underline{e}_2(\theta,\delta)$ are Gaussian as well, which supports an exact closed-form expression for the likelihood function of the raw residuals. Yet, SEP innovations, $\varepsilon(\theta,\delta) \sim \mathcal{SEP}(0,\sigma^2_\varepsilon,\beta,\xi)$, do not admit a closed-form description for $p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)$ and $p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)$ as the AR(2) process of Eq. (12) does not preserve higher-order moments of the SEP distribution such as its skew and kurtosis (Damsleth and El-Shaarawi, 1989; Hürlimann, 2012). We may resort to the law of total cumulance to determine the expected skew and kurtosis of $\underline{e}_1(\theta,\delta)$ and $\underline{e}_2(\theta,\delta)$, yet, such estimates must be turned into values of $\beta$ and $\xi$ to warrant a closed-form description of $p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)$ and $p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)$, respectively. Henceforth, we follow (Schoups and Vrugt, 2010) and condition on the unobserved residuals, $e_{-1}(\theta)$ and $e_0(\theta)$, to yield

$$p(\underline{e}_1|\underline{e}_0,\underline{e}_{-1};\theta,\delta)p(\underline{e}_2|\underline{e}_1,\underline{e}_0;\theta,\delta)$$

[2] In statistics, a studentized residual is equal to a residual divided by an estimate of its standard deviation. If instead we divide by the population standard deviation then we speak of a standardized residual

$$\simeq p(\underline{e}_1, \underline{e}_2 | \underline{e}_{-1}, \underline{e}_0; \boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{t=1}^{2} p(\underline{e}_t | \underline{e}_{t-1}, \underline{e}_{t-2}; \boldsymbol{\theta}, \boldsymbol{\delta}), \tag{17}$$

This approximation is valid only for moderate to large training data records, $\widetilde{\mathbf{y}}$. If we now substitute Eq. (17) into (16) and enter the resulting expression into Eq. (15), we yield the following expression for the conditional likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\delta}) &\simeq \left| \frac{\partial \underline{\mathbf{e}}(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \mathbf{e}(\boldsymbol{\theta})} \right| \prod_{t=1}^{n} p(\underline{e}_t | \underline{e}_{t-1}, \underline{e}_{t-2}; \boldsymbol{\theta}, \boldsymbol{\delta}) \\ &\simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} p(\underline{e}_t | \underline{e}_{t-1}, \underline{e}_{t-2}; \boldsymbol{\theta}, \boldsymbol{\delta}). \end{aligned} \tag{18}$$

According to the AR(2) process of Eq. (12), the conditional distribution of $\underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta})$ given $\underline{e}_{t-1}(\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\underline{e}_{t-1}(\boldsymbol{\theta}, \boldsymbol{\delta})$ equals the marginal distribution of the partial residuals, $\epsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta})$ for all $t = (1, 2, \dots, n)$. Hence, we can now write

$$L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\delta}) \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} f_{\text{SEP}}(\epsilon_t | 0, \sigma_\epsilon^2, \beta, \xi). \tag{19}$$

This formulation poses computational difficulties as we do not have available a mathematical expression for the non-standardized SEP distribution with variance, $\sigma_\varepsilon^2 \neq 1$. Fortunately, we can take advantage of the following identity

$$f_{\text{SEP}}(a | 0, \sigma_a^2, \beta, \xi) = \frac{1}{\sigma_a} f_{\text{SEP}}(\underline{a} | 0, 1, \beta, \xi), \tag{20}$$

and, thus, the non-standardized SEP-density, $f_{\text{SEP}}(a|0, \sigma_a^2, \beta, \xi)$, at point, $a$, with, $\text{Var}(a) = \sigma_a^2$, is equivalent to a multiple, $\sigma_a^{-1}$, of the standardized SEP-density, $f_{\text{SEP}}(\underline{a}|0, 1, \beta, \xi)$, evaluated at the standardized $a$-value, $\underline{a} = a/\sigma_a$. If we enter Eq. (20) into Eq. (19) then we end up with the following general formulation for the SEP likelihood function of the raw residuals

$$L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \boldsymbol{\delta}) \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} \frac{1}{\sigma_\varepsilon} f_{\text{SEP}}(\underline{\varepsilon}_t | 0, 1, \beta, \xi), \tag{21}$$

using standardized partial residuals, $\underline{\varepsilon}_t = \varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta})/\sigma_\varepsilon$ for all $t = (1, 2, \dots, n)$. We purposely use the terminology *standardized partial residual* as the denominator, $\sigma_\varepsilon$, equals the theoretical standard deviation of the AR(2) innovations in Eq. (13). If we admit the standardized SEP-density, $f_{\text{SEP}}\big(\cdot|0, 1, \beta, \xi\big)$, of Eq. (9) to Eq. (21) then we arrive at the SEP likelihood function

$$\begin{aligned} &L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \beta, \xi, \boldsymbol{\Phi}_2, \boldsymbol{\sigma}_\epsilon^2) \\ &\quad \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} \frac{1}{\sigma_\varepsilon} \frac{2\sigma_\xi \omega_\beta}{(\xi + \xi^{-1})} \exp\left( -c_\beta \left| \frac{\mu_\xi + \sigma_\xi \underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta})}{\xi^{\text{sign}(\mu_\xi + \sigma_\xi \underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}))}} \right|^{2/(1+\beta)} \right), \end{aligned} \tag{22}$$

where $c_\beta$, $\omega_\beta$, $\mu_\xi$, and $\sigma_\xi$ are a function of the kurtosis, $\beta$, and skewness, $\xi$, according to Eqs. (A.2a), (A.2b), (A.7a) and (A.7b), respectively, $\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta})$ signifies the $t^{\text{th}}$ standardized partial residual, $\sigma_\varepsilon^2$ is the variance of the AR(2)-process of Eq. (12), and $\boldsymbol{\delta} = [\, \beta \;\, \xi \;\, \boldsymbol{\Phi}_2 \,]^\top$. It is important to keep in mind that the SEP likelihood function is not exact but conditional upon the initial conditions, $e_{-1}(\boldsymbol{\theta})$ and $e_0(\boldsymbol{\theta})$.

Eq. (22) (but in modified form) was coined the *generalized likelihood (GL) function* by Schoups and Vrugt (2010) as it extends applicability to situations wherein the raw residuals, $\mathbf{e}(\boldsymbol{\theta})$, deviate from normality and exhibit skew, kurtosis, heteroscedasticity and, possibly, serial correlation at one or more lags. If as in Eq. (10) of Schoups and Vrugt (2010) the treatment of serial correlation precedes residual studentization then the multiplicative term, $1/\sigma_\varepsilon$, will disappear from Eq. (22). This original implementation is herein referred to as the GL function and the revised formulation of the SEP likelihood function in Eq. (22) is coined the GL$^+$ function.

Finite multiplication of the $n$ likelihoods in Eq. (22) is susceptible to arithmetic underflow, resulting in a zero value of $L(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \beta, \xi, \boldsymbol{\Phi}_2, \boldsymbol{\sigma}_\epsilon^2)$. For reasons of numerical stability it is therefore more convenient to work with the SEP log-likelihood function, $\mathcal{L}(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \beta, \xi, \boldsymbol{\Phi}_2, \boldsymbol{\sigma}_\epsilon^2)$, instead

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \beta, \xi, \boldsymbol{\Phi}_2, \boldsymbol{\sigma}_\epsilon^2) &\simeq n \log(2\sigma_\xi \omega_\beta) - n \log(\xi + \xi^{-1}) - \frac{1}{2} \sum_{t=1}^{n} \big\{ \log(\sigma_{\epsilon_t}^2) \big\} \\ &\quad - \frac{n}{2} \log(\sigma_\varepsilon^2) - c_\beta \sum_{t=1}^{n} \left| \frac{\mu_\xi + \sigma_\xi \underline{\epsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta})}{\xi^{\text{sign}(\mu_\xi + \sigma_\xi \underline{\epsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}))}} \right|^{2/(1+\beta)}, \end{aligned} \tag{23}$$

where $\sigma_\varepsilon^2$ is equal to Eq. (13) for the AR(2) process of the standardized raw residuals.

### 3.4. Treatment of nuisance variables and measurement data errors

The SEP log-likelihood function, $\mathcal{L}(\boldsymbol{\theta}|\widetilde{\mathbf{y}}, \beta, \xi, \boldsymbol{\Phi}_2, \boldsymbol{\sigma}_\epsilon^2)$, above assumes prior knowledge of the $s = 4$ nuisance variables, $\boldsymbol{\delta} = [\, \beta \;\, \xi \;\, \boldsymbol{\Phi}_2 \,]^\top$, and the $n$-vector of measurement error variances, $\boldsymbol{\sigma}_\epsilon^2 = [\, \sigma_{\epsilon_1}^2 \;\, \sigma_{\epsilon_2}^2 \;\, \dots \;\, \sigma_{\epsilon_n}^2 \,]^\top$, of the training data, $\widetilde{\mathbf{y}} = [\, \widetilde{y}_1 \;\, \widetilde{y}_2 \;\, \dots \;\, \widetilde{y}_n \,]^\top$. In the absence of detailed knowledge about the expected distribution of the residuals and/or measurement errors of $\widetilde{\mathbf{y}}$, we can choose to estimate the nuisance variables, $\boldsymbol{\delta}$, along with the model parameters, $\boldsymbol{\theta}$. This approach does not change anything to the mathematical formulation of the GL$^+$ function in Eq. (23), except the nuisance variables migrate to the left side of the vertical bar in the parent definition of the SEP log-likelihood function to read instead, $\mathcal{L}(\boldsymbol{\theta}, \beta, \xi, \boldsymbol{\Phi}_2|\widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2)$ or $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\delta}|\widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2)$. This alternative definition frees us from having to make strong (and often questionable) prior assumptions about the distribution of the residuals. The nuisance variables let the residuals speak for themselves. Note that nuisance variables are also referred to as hyperparameters when they relate to the prior distribution of the model parameters.

We can further generalize the SEP log-likelihood function of Eq. (23) to situations without knowledge of the training data measurement errors, $\boldsymbol{\sigma}_\epsilon^2$. A pragmatic remedy is to relate the measurement error to the simulated data (Schoups and Vrugt, 2010; Evin et al., 2013), as follows

$$s_{\epsilon_t} = s_0 + s_1 y_t(\boldsymbol{\theta}) \tag{24}$$

where $s_{\epsilon_t}$ is an estimate of the measurement error standard deviation of the $t^{\text{th}}$ simulated output, $y_t(\boldsymbol{\theta})$, and $s_0 > 0$ and $s_1 \in [0, 1]$ are unknown coefficients that define the intercept and slope of the measurement error function, respectively. With $s_1 = 0$, the expression above reduces to a homoscedastic measurement error function with variance equal to $s_0^2$.

In the original GL function of Schoups and Vrugt (2010), the slope and intercept of Eq. (24) were treated as two separate nuisance variables each having their own prior distribution. This implementation would be fine with the application of Eqs. (12) and (13) so that the variance of the partial residuals, $\sigma_\epsilon^2$, is directly proportional to the variance of the standardized raw residuals, $\sigma_{\underline{e}}^2$. In the GL$^+$ function, we simplified the implementation to include only a single nuisance variable, $s_0$. The slope, $s_1$, is evoked as a phantom variable to enforce a unit variance, $\sigma_{\underline{e}}^2 = 1$, of the studentized raw residuals, $\underline{\mathbf{e}}(\boldsymbol{\theta}, \boldsymbol{\delta})$. If $m_{\underline{e}}$ and $s_{\underline{e}}^2$ denote the sample mean and sample variance of the studentized raw residuals, respectively

$$m_{\underline{e}} = \frac{1}{n} \sum_{t=1}^{n} \underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) \qquad \text{and} \qquad s_{\underline{e}}^2 = \frac{1}{n-1} \sum_{t=1}^{n} \big( \underline{e}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) - m_{\underline{e}} \big)^2, \tag{25}$$

then the zero-point of the so-called residual function, $\hbar(s_1|\boldsymbol{\theta}, \boldsymbol{\delta}) = 1 - s_{\underline{e}}^2$, equals the value of $s_1$ for which $s_{\underline{e}}^2 = 1$. The secant method can efficiently solve this root-finding problem in only a handful of iterations. Thus, for a given value of the intercept, $s_0$, of Eq. (24) the secant method finds the value of the slope, $s_1$, so that $\sigma_{\underline{e}}^2 \rightarrow 1$. This approach increases the number of nuisance variables with only one, namely, $\boldsymbol{\delta} = [\, \beta \;\, \xi \;\, \boldsymbol{\Phi}_2 \;\, s_0 \,]^\top$, and the SEP log-likelihood function becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \beta, \xi, \boldsymbol{\Phi}_2, s_0|\widetilde{\mathbf{y}}) &\simeq n \log(2\sigma_\xi \omega_\beta) - n \log(\xi + \xi^{-1}) \\ &\quad - \sum_{t=1}^{n} \big\{ \log\big( |s_0 + s_1 y_t(\boldsymbol{\theta})| \big) \big\} \\ &\quad - \frac{n}{2} \log(\sigma_\varepsilon^2) - c_\beta \sum_{t=1}^{n} \left| \frac{\mu_\xi + \sigma_\xi \underline{\epsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta})}{\xi^{\text{sign}(\mu_\xi + \sigma_\xi \underline{\epsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}))}} \right|^{2/(1+\beta)}, \end{aligned} \tag{26}$$

**Table 1**
Nuisance variables of the GL$^+$ function, including their description, symbols, units, prior ranges and default values.

| Description | Symbol | Units | Min. | Max. | Default |
|---|---|---|---|---|---|
| Intercept of measurement error function | $s_0$ | mm/d | 0 | $\mathbb{R}_+$ | 0.1 |
| Kurtosis | $\beta$ | – | −1 | 1 | 0 |
| Skewness | $\xi$ | – | 0.1 | 10 | 1 |
| Autoregressive coefficient 1 | $\phi_1$ | – | 0 | 1 | 0 |
| Autoregressive coefficient 2 | $\phi_2$ | – | 0 | 1 | 0 |
| Slope of measurement error function | $s_1$ | – | 0 | $\mathbb{R}_+$ | 0 |

As our result will demonstrate our definition of $s_1$ as phantom variable rather than nuisance variable is inconsequential. Both methods produce equivalent results.

The root finding procedure is not limited to measurement errors with a non-constant variance, but could, in principle, also work for homoscedastic measurement errors. The value of $s_0$ will then determine whether the zero-point of $\hbar(s_1|\boldsymbol{\theta},\boldsymbol{\delta})$ is feasible or not. A simpler solution in this case, that avoids using any nuisance variable for the measurement error function of Eq. (24) after all, is to set $s_0$ equal to the sample standard deviation of the raw residuals, that is, $s_0 = \sqrt{s^2_e}$.

The inference of the slope and intercept of the measurement error function, is, however, not without problems. If the model is not able to describe sufficiently closely the observed data, $\widetilde{\mathbf{y}}$, then joint inference of $s_0$ and $\phi_1$ ($s_1$ is conditioned on $s_0$ via the constraint that $\sigma^2_{\underline{e}} \to 1$), may result in unrealistically large values of $s_0$ and/or $s_1$ while simultaneously promoting $\phi_1 \to 1$ so as to downplay the role of the model, $\mathcal{M}(\boldsymbol{\theta})$, in describing the measured data and maximize the SEP log-likelihood of Eq. (26). Unfortunately, there is nothing we can do about this. If the raw residuals, $\mathbf{e}(\boldsymbol{\theta})$, are large in magnitude then $s_0$ and/or $s_1$ must be large as well so as to honor the unit variance of the studentized raw residuals, $\underline{\mathbf{e}}(\boldsymbol{\theta},\boldsymbol{\delta})$. Instead, we must find ways to remove model bias and improve the overall model-data consistency. Forcing data errors will certainly contribute significantly to the model bias and simulation uncertainty.

Table 1 summarizes the nuisance variables of the GL$^+$ function and lists their lower and upper bounds and default values. If the measurement error variances, $\boldsymbol{\sigma}^2_\epsilon$, are known *a-priori* then the SEP log-likelihood function with AR(2) description of the standardized raw residuals, $\mathcal{L}(\boldsymbol{\theta},\beta,\xi,\boldsymbol{\Phi}_2|\widetilde{\mathbf{y}},\boldsymbol{\sigma}^2_\epsilon)$, is equal to Eq. (23) with $\boldsymbol{\Phi}_2 = [\,\phi_1\ \phi_2\,]^\top$ and $\sigma^2_\varepsilon$ equal to Eq. (13). If the $n$ measurement error variances, $\boldsymbol{\sigma}^2_\epsilon$, of the training data record, $\widetilde{\mathbf{y}}$, are unknown then we can resort to Eq. (26). This formulation, $\mathcal{L}(\boldsymbol{\theta},\beta,\xi,\boldsymbol{\Phi}_2,s_0|\widetilde{\mathbf{y}})$, requires specification of the measurement error intercept, $s_0$, in the case of heteroscedastic errors. This increases the number of nuisance variables to five, $\boldsymbol{\delta} = [\,\beta\ \xi\ \phi_1\ \phi_2\ s_0\,]^\top$. If an AR(1) is assumed instead, then $\phi_2 = 0$ and $\sigma^2_\varepsilon = 1 - \phi^2_1$ in Eqs. (23) and (26).

### 3.5. Special cases of the SEP log-likelihood function

If so desired, we can fix the skew, $\beta$, and kurtosis, $\xi$, of the SEP log-likelihood function to some default values to satisfy prior assumptions. For $\beta = 0$ and $\xi = 1$, Eq. (26) simplifies to the well-known Gaussian log-likelihood, $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\Phi}_2,s_0|\widetilde{\mathbf{y}},0,1)$, of an AR(2) process. Indeed, without skew and excess kurtosis, Eqs. (A.2a), (A.2b), (A.7a) and (A.7b), result in $c_\beta = \frac{1}{2}$, $\omega_\beta = (2\pi)^{-1/2}$, $\mu_\xi = 0$ and $\sigma_\xi = 1$. If we enter these values into the SEP log-likelihood function we yield

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\Phi}_2,s_0|\widetilde{\mathbf{y}},\beta=0,\xi=1)$$
$$\simeq -\frac{n}{2}\log(2\pi) - \sum_{t=1}^{n}\left\{\log\left(|s_0 + s_1 y_t(\boldsymbol{\theta})|\right)\right\} - \frac{n}{2}\log(\sigma^2_\varepsilon) - \frac{1}{2}\sum_{t=1}^{n}\underline{\epsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2, \tag{27}$$

which is equal to a conditional Gaussian log-likelihood function with constant and/or nonconstant measurement errors and AR(2)-process of the studentized raw residuals. This expression follows directly from the GLS likelihood function of Eq. (6). For $\phi_1 = 0$ and $\phi_2 = 0$, the above expression simplifies further to a normal log-likelihood function conditional upon the unobserved raw residuals, $e_{-1}(\boldsymbol{\theta})$ and $e_0(\boldsymbol{\theta})$, and (non)constant variance of the raw residuals. In the remainder of this paper, the normal log-likelihood of Eq. (27) will be referred to as NL function. For $\beta = 1$ and $\xi = 1$, the SEP-log likelihood function simplifies to

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\Phi}_2,s_0|\widetilde{\mathbf{y}},\beta=1,\xi=1)$$
$$\simeq -\frac{n}{2}\log(2) - \sum_{t=1}^{n}\left\{\log\left(|s_0 + s_1 y_t(\boldsymbol{\theta})|\right)\right\} - \frac{n}{2}\log(\sigma^2_\varepsilon) - \sqrt{2}\sum_{t=1}^{n}|\underline{\epsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})|, \tag{28}$$

which is equal to a Laplacian log-likelihood, $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\Phi}_2,s_0|\widetilde{\mathbf{y}},\beta=1,\xi=1)$, of an AR(2) process. This concludes our description of the GL$^+$ function.

### 3.6. The skewed Student's *t*-likelihood function

The GL and GL$^+$ functions rely on the generalized normal (or EP) distribution to better characterize nontraditional residual distributions with skew, kurtosis and heavier tails than the normal distribution. Scharnagl et al. (2015) replaced the EP distribution of Eq. (7) with a standardized Student's $t$-density

$$f_{\mathrm{ST}}(a|0,1,\nu) = \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi(\nu-2)}}\left(1+\frac{a^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{29}$$

where $\nu > 2$ denotes the degrees of freedom. If we combine Eq. (29) with the skew density of Eq. (8) then we yield the following expression for the standardized skewed Student's $t$ (SST) density (Scharnagl et al., 2015)

$$f_{\mathrm{SST}}(a|0,1,\nu,\xi)$$
$$= \frac{2\sigma_\xi}{(\xi+\xi^{-1})}\frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi a}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi a)}}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{30}$$

with skewness parameter, $\xi > 0$, and shift and scale constants, $\mu_\xi$ and $\sigma^2_\xi$, respectively, that depend on $\nu$ and $\xi$ and standardize the SST density (Scharnagl et al., 2015). Closed-form expressions for $\mu_\xi$ and $\sigma^2_\xi$ are presented in Eqs. (B.7a) and (B.7b) of Appendix B.

Fig. 2 plots the density of the standardized skewed Student $t$-distribution of Eq. (30) for $a \in [-3\frac{1}{2}, 3\frac{1}{2}]$ using different values of the (a) skewness, $\xi$, and (b) degrees of freedom, $\nu$.

The SST density is positively skewed for $\xi > 1$ and negatively skewed for $\xi < 1$. For $\xi = 1$, the Student's $t$-distribution is symmetric and bell-shaped albeit with (much) heavier tails than the normal distribution for small values of $\nu$. These so-called leptokurtic distributions are more peaked than a normal distribution and have stronger tails. For large values of $\nu$ the Student's $t$-distribution converges to the standard normal distribution (zero mean and unit standard deviation). Thus, the value of $\nu \in (2,\infty)$ determines the kurtosis and heaviness of the tails of the SST distribution.

If we assume an AR(2)-process of the standardized raw residuals and treat measurement data errors as in Section 3.4 then we yield the log-likelihood function, $\mathcal{L}(\boldsymbol{\theta},s_0,\nu,\xi,\sigma^2_\varepsilon|\widetilde{\mathbf{y}})$, of the SST density of Scharnagl et al. (2015) (see Appendix B)

$$\mathcal{L}(\boldsymbol{\theta},s_0,\nu,\xi,\boldsymbol{\Phi}_2|\widetilde{\mathbf{y}})$$

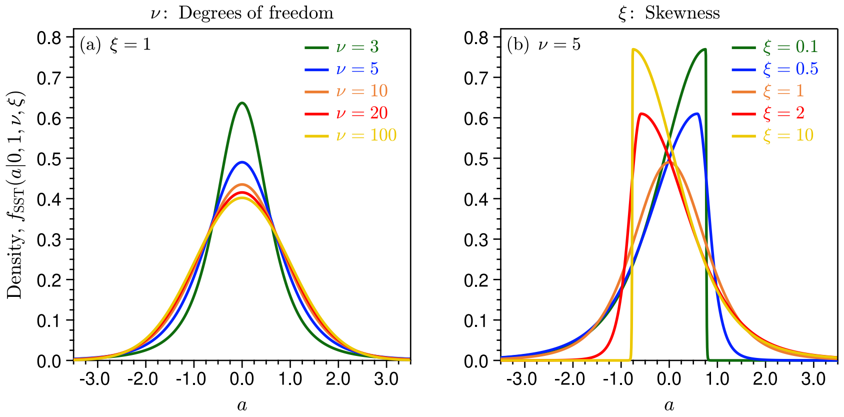


**Fig. 2.** Density of the standardized skewed Student's $t$-distribution of Eq. (30) for $\xi = 1$ and $\nu = 5$ degrees of freedom. Color coding differentiates between different values of $\nu$ (left panel) and $\xi$ (right panel). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Nuisance variables of the SL function, including their description, symbols, units, prior ranges and default values.

| Description | Symbol | Units | Min. | Max. | Default |
|---|---|---|---|---|---|
| Intercept of measurement error function | $s_0$ | mm/d | 0 | $\mathbb{R}_+$ | 0.1 |
| Degrees of freedom | $\nu$ | – | 2 | $\infty$ | $n-d$ |
| Skewness | $\xi$ | – | 0.1 | 10 | 1 |
| Autoregressive coefficient 1 | $\phi_1$ | – | 0 | 1 | 0 |
| Autoregressive coefficient 2 | $\phi_2$ | – | 0 | 1 | 0 |
| Slope of measurement error function | $s_1$ | – | 0 | $\mathbb{R}_+$ | 0 |

$$\begin{aligned}&= -\frac{n}{2}\log(\sigma_\epsilon^2) - \sum_{t=1}^{n}\left\{\log\big(|s_0 + s_1 y_t(\boldsymbol{\theta})|\big)\right\} + n\log(2) + n\log(\sigma_\xi)\\ &\quad + n\log\Big(\Gamma\big((\nu+1)/2\big)\Big) - n\log(\xi+\xi^{-1}) - n\log\big(\Gamma(\nu/2)\big) - \frac{n}{2}\log(\pi)\\ &\quad - \frac{n}{2}\log(\nu-2) - \frac{\nu+1}{2}\sum_{t=1}^{n}\left\{\log\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi + \sigma_\xi\varepsilon_t(\boldsymbol{\theta},\boldsymbol{\delta})}{\xi^{\text{sign}(\mu_\xi+\sigma_\xi\varepsilon_t(\boldsymbol{\theta},\boldsymbol{\delta}))}}\right)^2\right)\right\},\end{aligned} \tag{31}$$

where $s_1$ equals the zero-point of the residual function, $\hbar(s_1|\boldsymbol{\theta},\boldsymbol{\delta}) = 1 - s_e^2$, so as to enforce a unit variance of the studentized raw residuals. The SST log-likelihood is referred to as Student likelihood or SL function in the remainder of this paper.

Table 2 summarizes the nuisance variables of the SL function. The default value of the degrees of freedom, $\nu$, is set equal to the sample size, $n$, minus the number, $d$, of estimated parameters, $\boldsymbol{\theta}$. Note that for $\xi = 1$ and $\nu \to \infty$ the SL function reduces to the NL function of Eq. (27).

## 4. The universal likelihood function

The skewed generalized Student's $t$-(SGT)-distribution of Theodossiou (1998) is a further generalization of the SST distribution in Eq. (30) to a family of continuous probability distributions, which includes the generalized Student's $t$ (McDonald and Newey, 1988; Hansen, 1994), error (Theodossiou, 2015) and exponential power distributions (Subbotin, 1923; Box and Tiao, 1992), the skewed and symmetric Laplace, Cauchy, Student's $t$ and normal distributions (see e.g. Johnson et al. (1995)) and the uniform distribution. The PDF of the non-standardized SGT distribution is given by Theodossiou (2015)

$$\begin{aligned}&f_{\text{SGT}}(a|\mu,\sigma,\lambda,p,q)\\ &= \frac{p}{2\kappa_{\lambda pq}\sigma B\big(1/p,q/p\big)}\left(1+\left|\frac{a-\mu+\mu_{\lambda pq}}{\kappa_{\lambda pq}\sigma\big(1+\lambda\,\text{sign}(a-\mu+\mu_{\lambda pq})\big)}\right|^p\right)^{-(q+1)/p},\end{aligned} \tag{32}$$

where $\mu$, $\sigma > 0$ and $\lambda \in (-1,1)$ are location, scale and skewness parameters, $p > 0$ and $q > 0$ control the kurtosis of the distribution, $\mu_{\lambda pq}$ and $\kappa_{\lambda pq}$, are shift and scale constants, respectively, that depend on $\lambda$, $p$ and $q$, and $B(a,b)$ is the so-called beta function or Euler integral of the first kind

$$B(a,b) = \int_0^1 x^{(a-1)}(1-x)^{(b-1)}\,\mathrm{d}x, \tag{33}$$

and is equivalent to the following expression

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{34}$$

using the gamma function, $\Gamma(\cdot)$, in Eq. (A.3).

The scalars $\mu_{\lambda pq}$ and $\kappa_{\lambda pq}$ negate changes in the mean, $\mu$, and variance, $\sigma^2$, of the SGT distribution imposed by the variables, $p$, $q$ and $\lambda$. In the original parameterization of Theodossiou (1998), the shift and scale constants are defined as given in Eqs. (35a) and (35b) (see Box I) and yield a SGT distribution with mean, $\mu$, and variance, $\sigma^2$, if $pq > 2$. If we admit, $\mu = 0$ and $\sigma^2 = 1$, then we yield the standardized SGT density

$$\begin{aligned}&f_{\text{SGT}}(a|0,1,\lambda,p,q)\\ &= \frac{p}{2\kappa_{\lambda pq}B\big(1/p,q/p\big)}\left(1+\left|\frac{a+\mu_{\lambda pq}}{\kappa_{\lambda pq}\big(1+\lambda\,\text{sign}(a+\mu_{\lambda pq})\big)}\right|^p\right)^{-(q+1)/p}.\end{aligned} \tag{36}$$

The SGT distribution of Eq. (36) with the closed-form expressions for $\mu_{\lambda pq}$ and $\kappa_{\lambda pq}$ in Eqs. (35a) and (35b) nests a number of commonly used univariate probability distributions including (but not limited to) the generalized Student's $t$-distribution (McDonald and Newey, 1988), the generalized error distribution[3], and the symmetric and skewed Student's $t$, Laplace, Cauchy-Lorentz and normal distributions (see Table 3 and Appendix C).

To provide insights into the functional form of Eq. (36), Fig. 3 displays the standardized SGT density, $f_{\text{SGT}}(a|0,1,\lambda,p,q)$, for $-3\frac{1}{2} \leq a \leq 3\frac{1}{2}$ for different values of the (a) kurtosis, $p > 0$, (b) skewness, $-1 < \lambda < 1$, (c) kurtosis, $q > 0$, and (d) $p$, $\beta$ and $q$.

The SGT distribution can be symmetric and highly skewed. For $\lambda \in (-1,0)$, the distribution is negatively skewed (to the left), and for $\lambda \in (0,1)$, the SGT distribution is positively skewed (to the right). For

[3] Also known as the generalized normal distribution or EP distribution

$$\mu_{\lambda pq} = \frac{2\kappa_{\lambda pq}\lambda\sigma B\big(2/p,(q-1)/p\big)}{B(1/p,q/p)} \qquad \text{if } pq > 1 \tag{35a}$$

$$\kappa_{\lambda pq} = \frac{B(1/p,q/p)}{\sqrt{(1+3\lambda^2)B(1/p,q/p)B\big(3/p,(q-2)/p\big)-4\lambda^2 B\big(2/p,(q-1)/p\big)^2}} \qquad \text{if } pq > 2, \tag{35b}$$

**Box I.**

**Table 3**
Special and/or limiting cases of the Skewed Generalized Student's *t*-distribution.

| $\lambda$ | $p$ | $q$ | Distribution | Reference |
|---|---|---|---|---|
| 0 | | | Generalized Student's *t*-distribution | McDonald and Newey (1988) |
| | 2 | | Skewed Student's *t*-distribution | Hansen (1994) |
| 0 | 2 | | Student's *t* distribution | Student (1908) |
| | | $\infty$ | Skewed exponential power distribution | ‡ |
| 0 | | $\infty$ | Exponential power distribution | Subbotin (1923), Box and Tiao (1992) |
| | 1 | $\infty$ | Skewed Laplace distribution | |
| 0 | 1 | $\infty$ | Laplace distribution | Laplace (1774) |
| | 2 | 1 | Skewed Cauchy–Lorentz distribution | |
| 0 | 2 | 1 | Cauchy–Lorentz distribution | Poisson (1824) |
| | 2 | $\infty$ | Skewed normal distribution | |
| 0 | 2 | $\infty$ | Normal distribution | |
| 0 | $\infty$ | $\infty$ | Uniform distribution | |

‡Treatment of skew is different than in template density of Fernandez and Steel (1998).

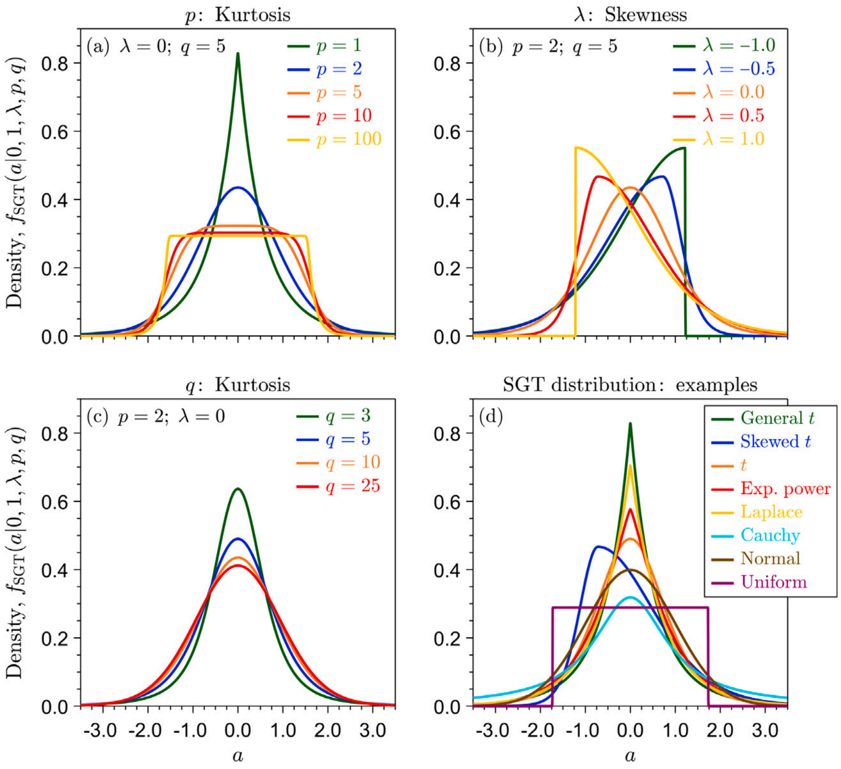


**Fig. 3.** Probability density function of the standardized skewed generalized *t*-distribution, $f_{\mathrm{SGT}}(a|0,1,\lambda,p,q)$, of Eq. (36) for different values of (a) the kurtosis, $p$, (b) skewness, $\lambda$, (c) kurtosis, $q$, and (d) $p$, $q$ and $\lambda$ of the special SGT parameterizations listed in Table 3. The coefficients, $\mu_{\lambda pq}$ and $\kappa_{\lambda pq}$, are computed using the closed-form expressions of Eqs. (35a) and (35b), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$\lambda = 0$, the distribution is symmetric. The parameters $p > 0$ and $q > 2$ control the kurtosis of the SGT distribution. The smaller the values of $p$ and $q$, the more slender the SGT distribution will be, and, thus, the larger its kurtosis. A distribution with kurtosis that exceeds the value of three is also called leptokurtic and will have fatter tails than a normal distribution. For large values of $p$ and $q$ the excess kurtosis of the SGT distribution disappears in lieu of a much more uniform distribution of the probability density. The parameter $p$ exerts the largest control on the kurtosis, as is evident by the large variation in SGT densities for different values of $p$. The impact of $q$ on the SGT distribution is more subtle, and its impact decreases substantially with large values of $q$. Altogether, small values of $p$ and/or $q$ produce a SGT distribution with much heavier tails than the normal density. The bottom panel demonstrates the ability of the standardized SGT density to describe a wide variety of distributions.

The graph plots four different parameterizations (color red) of the standardized SGT distribution. Note, that in each graph, the mean of the SGT distribution equals zero — and, more difficult to see, the variance equals one. The bottom-right graph presents eight of the special parameterizations of the SGT distribution listed in Table 3, among which the EP distribution of Eq. (7) (in red). The ability of the SGT distribution to morph into a wide variety of well-known probability distributions has desirable advantages for residual characterization and description. Henceforth, the forthcoming likelihood function stemming from the SGT distribution is referred to as the universal likelihood function.

### 4.1. Derivation of SGT (log)-likelihood function

If we treat the SGT parameters, $\lambda$, $p$ and $q$, as nuisance variables and assume further that the studentized raw residuals, $\underline{\mathbf{e}}(\theta, \delta)$, exhibit serial correlation according to the AR(2) scheme of Eqs. (12) then the SGT likelihood should follow our derivation of the SEP likelihood function. Specifically, according to Eq. (18) we may write

$$L(\boldsymbol{\theta}, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2) \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} p(\underline{e}_t | \underline{e}_{t-1}, \underline{e}_{t-2}; \boldsymbol{\theta}, \boldsymbol{\delta}) \tag{37}$$

and, thus, we yield

$$\begin{aligned} & L(\boldsymbol{\theta}, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2) \\ & \quad \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} f_{\text{SGT}}\big(\varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta}) | 0, \sigma_\varepsilon^2, \lambda, p, q\big) \\ & \quad \simeq \prod_{t=1}^{n} \frac{1}{\sigma_{\epsilon_t}} \frac{1}{\sigma_\varepsilon} f_{\text{SGT}}\big(\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) | 0, 1, \lambda, p, q\big) \\ & \quad \simeq \sigma_\varepsilon^{-n} \prod_{t=1}^{n} \frac{p}{2\sigma_{\epsilon_t} \kappa_{\lambda pq} B(1/p, q/p)} \\ & \quad \times \quad \left(1 + \left| \frac{\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq}}{\kappa_{\lambda pq}\big(1 + \lambda \operatorname{sign}(\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq})\big)} \right|^p \right)^{-(q+1)/p} \end{aligned} \tag{38}$$

where $f_{\text{SGT}}\big(\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) | 0, 1, \lambda, p, q\big)$ returns the density at $\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta})$ of the standardized SGT distribution in Eq. (36), $\underline{\boldsymbol{\varepsilon}}(\boldsymbol{\theta}, \boldsymbol{\delta}) = [\, \underline{\varepsilon}_1(\boldsymbol{\theta}, \boldsymbol{\delta}) \; \underline{\varepsilon}_2(\boldsymbol{\theta}, \boldsymbol{\delta}) \; \ldots \; \underline{\varepsilon}_n(\boldsymbol{\theta}, \boldsymbol{\delta}) \,]^\top$, denotes the $n$-vector of standardized partial residuals and $\sigma_\varepsilon^2$ is the variance of the partial residuals, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\delta}) = [\, \varepsilon_1(\boldsymbol{\theta}, \boldsymbol{\delta}) \; \varepsilon_2(\boldsymbol{\theta}, \boldsymbol{\delta}) \; \ldots \; \varepsilon_n(\boldsymbol{\theta}, \boldsymbol{\delta}) \,]^\top$, computed using Eq. (13). If we take the natural logarithm of Eq. (38) then we yield the SGT log-likelihood function

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\theta}, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2) \\ & \quad \simeq -\frac{n}{2} \log(\sigma_\varepsilon^2) - \sum_{t=1}^{n} \big\{\log(\sigma_{\epsilon_t})\big\} + n \log(p) \\ & \qquad - n \log(2) - n \log(\kappa_{\lambda pq}) - n \log\big(B(1/p, q/p)\big) \\ & \qquad - \frac{q+1}{p} \sum_{t=1}^{n} \left\{ \log\left( 1 + \left| \frac{\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq}}{\kappa_{\lambda pq}\big(1 + \lambda \operatorname{sign}(\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq})\big)} \right|^p \right) \right\}, \end{aligned} \tag{39}$$

where the natural logarithm of the beta function, $B(a, b)$, follows from Eq. (34)

$$\log\big(B(a, b)\big) = \log\big(\Gamma(a)\big) + \log\big(\Gamma(b)\big) - \log\big(\Gamma(a - b)\big). \tag{40}$$

The SGT (log)-likelihood functions in Eqs. (38) and (39) assume prior knowledge of the measurement error variances, $\boldsymbol{\sigma}_\epsilon^2$, of the training data record, $\widetilde{\mathbf{y}}$. If the $n$ entries of $\boldsymbol{\sigma}_\epsilon^2$ are unknown then we can resort to the measurement error function of Eq. (24) and treat the intercept, $s_0$, as an additional nuisance variable. Then the SGT log-likelihood function becomes

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\theta}, s_0, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}}) \\ & \quad \simeq -\frac{n}{2} \log(\sigma_\varepsilon^2) - \sum_{t=1}^{n} \big\{\log\big(|s_0 + s_1 y_t(\boldsymbol{\theta})|\big)\big\} + n \log(p) \\ & \qquad - n \log(2) - n \log(\kappa_{\lambda pq}) - n \log\big(B(1/p, q/p)\big) \\ & \qquad - \frac{q+1}{p} \sum_{t=1}^{n} \left\{ \log\left( 1 + \left| \frac{\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq}}{\kappa_{\lambda pq}\big(1 + \lambda \operatorname{sign}(\underline{\varepsilon}_t(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mu_{\lambda pq})\big)} \right|^p \right) \right\}. \end{aligned} \tag{41}$$

where $s_1$ is computed using root finding of the residual function, $\hbar(s_1 | \boldsymbol{\theta}, \boldsymbol{\delta}) = 1 - s_{\underline{e}}^2$, with Newton's method. For the AR(2) process of the raw studentized residuals in Eq. (12), the variance, $\sigma_\varepsilon^2$, of the SGT innovations equals Eq. (13). If the second autoregressive coefficient, $\phi_2$, is fixed at zero then the variance of the partial residuals, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\delta})$, simplifies to $\sigma_\varepsilon^2 = \sigma_{\underline{e}}^2/(1 - \phi_1^2)$, the variance of the innovations of an AR(1) process. In the remainder of this paper, we refer to Eq. (41) as the universal likelihood (UL) function.

Table 4 summarizes the nuisance variables of the UL function including symbol, units, prior ranges and default values. Three remarks are in order. First, the product, $p \times q$, of the kurtosis parameters, must always exceed two, otherwise the scale constant, $\kappa_{\lambda pq}$, is not defined. Second, the upper bounds of $p$ and $q$ are arbitrary. They should just be set large enough so as to benefit from the versatility of the SGT distribution. Third, the default values of $\lambda = 0$, $p = 2$ and $q = 10^{10}$ imply a normal distribution (see Table 3).

### 4.2. Special and/or limiting cases of the SGT log-likelihood function

The two formulations of the UL function, $\mathcal{L}(\boldsymbol{\theta}, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}}, \boldsymbol{\sigma}_\epsilon^2)$ and $\mathcal{L}(\boldsymbol{\theta}, s_0, \lambda, p, q, \boldsymbol{\Phi}_2 | \widetilde{\mathbf{y}})$, assume joint inference of the skewness, $\lambda$, and kurtosis, $p$ and $q$, of the partial residuals and the model parameters, $\boldsymbol{\theta}$. If the partial residuals, $\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\delta})$, should follow some known distribution then we can set one or more nuisance variables at a default value. For example, if the $\varepsilon_t(\boldsymbol{\theta}, \boldsymbol{\delta})$'s are expected to follow a Student's $t$ distribution, then one can fix $\lambda = 0$ and $p = 2$ (see Fig. 3d and Table 3) and infer the remaining nuisance variables, $q$, $s_0$, and/or, $\phi_1$ and/or $\phi_2$, of the SGT log-likelihood function with the $d$ model parameters, $\boldsymbol{\theta}$. Note that in this case, the shift constant, $\mu_{\lambda pq}$, in Eq. (35a) equals $\mu_{02q} = 0$, the scale constant, $\kappa_{02q} = \sqrt{q - 2}$, and the SGT log-likelihood function of Eq. (41) reduces to the SST log-likelihood of Eq. (31) with $\xi = 1$. This is only one of many special cases of the SGT function as discussed in Appendix C. The more common case of normally distributed (partial) residuals deserves specific attention in Appendix D.

In summary, the SGT distribution is the key ingredient of a family of likelihood functions, which includes the normal likelihood function. Thus, the UL function makes obsolete other commonly used likelihood functions in DREAM Suite.

## 5. Case studies

In this section, we illustrate the application of the GL, GL$^+$, SL and UL functions to two case studies involving a simple autoregressive scheme and a conceptual watershed model. We first detail the methodological steps and then present the individual studies and their findings.

### 5.1. Posterior approximation

Unless we use a conjugate prior, the GL, GL$^+$, SL and UL functions, $L(\boldsymbol{\theta}, \boldsymbol{\delta} | \widetilde{\mathbf{y}}, \cdot)$, do not admit a closed-form solution for the posterior parameter distribution, $p(\boldsymbol{\theta}, \boldsymbol{\delta} | \widetilde{\mathbf{y}}, \cdot) \propto p(\boldsymbol{\theta}, \boldsymbol{\delta}) L(\boldsymbol{\theta}, \boldsymbol{\delta} | \widetilde{\mathbf{y}}, \cdot)$. We therefore resort to numerical integration (Gelfand and Smith, 1990; Gilks et al., 1996) and infer the joint posterior distribution of the $d$ model parameters, $\boldsymbol{\theta}$, and $s$ nuisance variables, $\boldsymbol{\delta}$, using Markov chain Monte Carlo simulation with DREAM Suite, a Windows-based implementation of the DiffeRential Evolution Adaptive Metropolis algorithm (Vrugt et al., 2008, 2009; Vrugt, 2016) and variants thereof Vrugt and ter Braak (2011), Laloy and Vrugt (2012), Vrugt and Sadegh (2013a), Sadegh et al. (2015), Vrugt and Beven (2018). The transition kernel of DREAM creates multiple different sequences (chains) of parameter vectors that are stationary

**Table 4**
Nuisance variables of the UL function, including their description, symbols, units, prior ranges and default values.

| Description | Symbol | Units | Min. | Max. | Default |
|---|---|---|---|---|---|
| Intercept of measurement error function | $s_0$ | mm/d | 0 | $\mathbb{R}_+$ | 0.1 |
| Skewness | $\lambda$ | – | −1 | 1 | 0 |
| Kurtosis | $p$ | – | 0 | 10 | 2 |
| Kurtosis | $q$ | – | 0 | 100 | $10^{10}$ |
| Autoregressive coefficient 1 | $\phi_1$ | – | 0 | 1 | 0 |
| Autoregressive coefficient 2 | $\phi_2$ | – | 0 | 1 | 0 |
| Slope of measurement error function | $s_1$ | – | 0 | $\mathbb{R}_+$ | 0 |

and ergodic and have as joint distribution, $p(\boldsymbol{\theta},\boldsymbol{\delta}|\widetilde{\mathbf{y}},\cdot)$ (ter Braak, 2006; Vrugt et al., 2009). Convergence of the joint chains to a stationary distribution is monitored using several different diagnostic measures as advocated by Cowles and Carlin (1996). This includes single chain and multi-chain diagnostics, among which the convergence metrics of Raftery and Lewis (1992) and Geweke (1992) and the univariate and multivariate scale-reduction factors, $\widehat{R}^d$ and $\widehat{R}$, of Gelman and Rubin (1992) and Brooks and Gelman (1998), respectively. The $m$ posterior vectors of the joint Markov chains are stored column-wise in a $(d+s)\times m$ matrix, $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$.

### 5.2. Performance of optimal parameter values

The maximum a posteriori (MAP) parameter values, $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}})$, equal the columns of $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$[4] which maximize the unnormalized posterior density, $p(\boldsymbol{\theta},\boldsymbol{\delta}|\widetilde{\mathbf{y}},\cdot)$. The performance of the MAP point predictions may be quantified using different summary measures, also known as scoring functions. We use the root mean square error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\widetilde{y}_t - y_t(\widehat{\boldsymbol{\theta}})\right)^2} \tag{42}$$

and percentage bias

$$\text{PBIAS} = 100\left(\frac{\sum_{t=1}^{n}\left(y_t(\widehat{\boldsymbol{\theta}}) - \widetilde{y}_t\right)}{\sum_{t=1}^{n}\widetilde{y}_t}\right). \tag{43}$$

Furthermore, we also perform diagnostic checks of the residuals of the MAP solution. Specifically, we investigate the homoscedasticity, distribution, autocorrelation function and quantiles of the standardized partial residuals, $\underline{\boldsymbol{\varepsilon}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}})$, of the MAP solution. The sample autocorrelation, $\rho_{\underline{\varepsilon}}(k)$, of two standardized partial residuals, $\underline{\varepsilon}_i(\boldsymbol{\theta},\boldsymbol{\delta})$ and $\underline{\varepsilon}_j(\boldsymbol{\theta},\boldsymbol{\delta})$, a distance (time) $k = |i-j|$ apart may be computed using

$$\begin{aligned}\widehat{\rho}_{\underline{\varepsilon}}(k) &= \frac{\text{Cov}\left(\underline{\varepsilon}_i(\boldsymbol{\theta},\boldsymbol{\delta}),\underline{\varepsilon}_j(\boldsymbol{\theta},\boldsymbol{\delta})\right)}{\text{Var}\left(\underline{\varepsilon}_i(\boldsymbol{\theta},\boldsymbol{\delta})\right)}\\ &= \frac{\frac{1}{n-k}\sum_{i=k+1}^{n}\left(\underline{\varepsilon}_i(\boldsymbol{\theta},\boldsymbol{\delta}) - m_{\underline{\varepsilon}}\right)\left(\underline{\varepsilon}_{i-k}(\boldsymbol{\theta},\boldsymbol{\delta}) - m_{\underline{\varepsilon}}\right)}{\frac{1}{n-k}\sum_{i=k+1}^{n}\left(\underline{\varepsilon}_i(\boldsymbol{\theta},\boldsymbol{\delta}) - m_{\underline{\varepsilon}}\right)^2},\end{aligned} \tag{44}$$

where $m_{\underline{\varepsilon}} = \frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})$ is the unitless mean of the $n$-record of standardized partial residuals and $k = (1,2,\ldots,n-1)$.

### 5.3. Parameter and predictive uncertainty

Parameter uncertainty is computed following Eq. (2) by truncating $\alpha/2$ of the left and right tails of the posterior samples, $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$, to yield $100(1-\alpha)\%$ credible intervals, $C_\alpha = (a,b)$, of the parameters and nuisance variables. Stationarity and ergodicity of the sampled chains of DREAM Suite imply that the predictions, $y_1^*, y_2^*, \ldots, y_m^*$, of the $m$ posterior parameter vectors, $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$, have as invariant distribution the posterior predictive PDF, $p(y|\widetilde{\mathbf{y}})$, and CDF, $F(y|\widetilde{\mathbf{y}})$, in Eqs. (3) and (4), respectively. The predictive distribution, $p(y|\widetilde{\mathbf{y}})$, of Eq. (3), can be obtained by marginalizing each model prediction over the posterior distribution. A detailed algorithmic recipe on how to do this is provided in Appendix E using the posterior realizations of the GL$^+$ function. This supplement also returns predictive percentiles, $\mathbf{y}^*_{t,\alpha/2}$ and $\mathbf{y}^*_{t,1-\alpha/2}$, which together define the $100(1-\alpha)\%$ prediction interval of $p(y_t|\widetilde{\mathbf{y}})$, where $t = (1,2,\ldots,n)$.

### 5.4. Comparison and ranking of predictive distributions

We use so-called scoring rules to evaluate the quality of the predictive distributions derived from the different likelihood functions. Scoring rules condense the accuracy of a distribution forecast to a single penalty oriented value while retaining attractive statistical properties (Alexander et al., 2022).

If $\Omega \subseteq \mathbb{R}$ denotes the set of possible values of the quantity of interest, $y$, and $\mathcal{F}$ is a convex class of probability distributions on $\Omega$ then a scoring rule, $\mathcal{S}(F,\tilde{y})$, is a function

$$\mathcal{S} : \mathcal{F}\times\Omega \longrightarrow \mathbb{R}, \tag{45}$$

that assigns numerical values to pairs of forecasts, $F \in \mathcal{F}$, and observations, $\widetilde{y} \in \Omega$.[5] Based on early recommendations by Brier (1950) and Shuford et al. (1966), we restrict attention to *proper* scoring rules, the statistical implications of which have been discussed in the literature by Gneiting and Raftery (2007). A scoring rule is *proper* relative to $\mathcal{F}$ if the expected score

$$\mathcal{S}(F,G) = \int_\Omega \mathcal{S}(F,\tilde{y})\mathrm{d}G(\widetilde{y}) \tag{46}$$

is minimized for $F = G$, and, thus, $\mathcal{S}(G,G) \leq \mathcal{S}(F,G)$ for all probability distributions, $F,G \in \mathcal{F}$ (Krüger et al., 2021). A score rule is *strictly proper* relative to the class $\mathcal{F}$ if the above holds with equality only if $F = G$ (Lerch et al., 2017). Such *strictly proper* scoring rules have an important advantage over *proper* scoring rules. In statistical jargon, a *strictly proper* score rule is a sufficient condition, whereas a *proper* score rule is a necessary but not sufficient condition. In plain words, if $\mathcal{S}(F,G)$ is a *strictly proper* score rule, then the smaller its value, the closer the distribution of $F$ will be to that of $G$. This is not true for *proper* scoring rules, which can attain a perfect score even if $F \neq G$.

While it is generally agreed upon that scoring rules must at least be proper to adequately quantify the accuracy of probabilistic forecasts (Winkler et al., 1996; Gneiting and Ranjan, 2011), the question which ones to use remains largely open (Gneiting and Raftery, 2007; Alexander et al., 2022). We restrict our attention to the four scoring rules listed in Table 5. These scoring rules benefit a strong mathematical underpinning (Dawid, 2007; Gneiting and Raftery, 2007) and are used to evaluate predictive distributions of MCMC output in probabilistic forecasting (Krüger et al., 2021). The scoring rules are

[4] The semicolon implies vertical concatenation. Thus, matrix $\boldsymbol{\Delta}^*$ is concatenated to the bottom of matrix $\boldsymbol{\Theta}^*$

[5] We use the terminology "forecasts" for the residual corrupted model output, $\mathbf{y}_j^* = \mathbf{y}_j(\boldsymbol{\theta}) + \mathbf{e}_j(\boldsymbol{\delta}) \in \mathbb{R}^{n\times m}$, of the posterior realizations, $j = (1,2,\ldots,m)$. Strictly speaking, these outcomes are not forecasts as the simulated output, $\mathbf{y}_j(\boldsymbol{\theta})$ assumes knowledge of the exogenous variables.

**Table 5**

Summary of the instantaneous scoring rules, $\mathcal{S}(F_t, \widetilde{y}_t)$, used to evaluate the quality of the predictive distributions of the different likelihood functions. For a probability distribution with CDF, $F_t$, we write $f_t$ for its density.

| Scoring rule | Symbol | $\mathcal{S}(F_t, \widetilde{y}_t)$ | Reference |
|---|---|---|---|
| Logarithmic[a] | LS | $-\log\bigl(f_t(\widetilde{y}_t)\bigr)$ | Good (1952) |
| Cont. rank probability[b] | CRPS | $\int_{-\infty}^{\infty} \bigl(F_t(z) - \mathbb{1}\{\widetilde{y}_t \leq z\}\bigr)^2 \mathrm{d}z$ | Matheson and Winkler (1976) |
| Pseudospherical[c] | PS | $-\dfrac{f_t(\widetilde{y}_t)^{\zeta-1}}{\Bigl(\int_{-\infty}^{\infty} f_t(z)^{\zeta}\mathrm{d}z\Bigr)^{(\zeta-1)/\zeta}}$ | Good (1971) |
| Interval[d] | IS | $(u_t - l_t) + \frac{2}{\alpha}(l_t - \widetilde{y}_t)\mathbb{1}\{\widetilde{y}_t < l_t\}$<br>$+\frac{2}{\alpha}(\widetilde{y}_t - u_t)\mathbb{1}\{\widetilde{y}_t > u_t\}$ | Gneiting and Raftery (2007) |

[a] Limiting case ($\zeta \rightarrow 1$) of the pseudospherical score if suitably scaled.
[b] The notation, $\mathbb{1}\{\widetilde{y}_t \leq z\}$, signifies the empirical CDF of the scalar observation, $\widetilde{y}_t$.
[c] The norm, $\zeta$, must exceed one and the minus sign reverses the orientation.
[d] Lower, $l_t = y^*_{t,\alpha/2}$, and upper, $u_t = y^*_{t,1-\alpha/2}$, limits of $100(1-\alpha)\%$ prediction interval.

**Table 6**

Time-averaged values of the performance metrics, $\mathcal{P}(\mathbb{F}, \widetilde{\mathbf{y}})$, of the predictive distributions derived from the $m$ posterior realizations.

| Performance metric | Symbol | $\mathcal{P}(\mathbb{F}, \widetilde{\mathbf{y}})$ | Reference |
|---|---|---|---|
| Reliability[a] | RLBL | $1 - \frac{2}{n}\sum_{j=1}^{n}\Bigl\lvert F^{\mathrm{s}}_j - \frac{j}{n}\Bigr\rvert$ | Renard et al. (2011) |
| Coefficient of variation | CV | $\frac{1}{n}\sum_{t=1}^{n}\dfrac{s_{F_t}}{m_{F_t}}$ | Evin et al. (2013) |
| Coverage | $C$ | $\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\{l_t \leq \widetilde{y}_t \leq u_t\}$ | |
| Width | $W$ | $\frac{1}{n}\sum_{t=1}^{n}(u_t - l_t)$ | |

[a] $F^{\mathrm{s}}_1 \leq F^{\mathrm{s}}_2 \leq \cdots \leq F^{\mathrm{s}}_n$; the sorted eCDFs, $F_t(\widetilde{y}_t)$, of (48) for all $t = (1, 2, \ldots, n)$.

negatively oriented,[6] and for all but the interval score (IS) condense different aspects of a forecast distribution among which accuracy, skill, resolution and sharpness. In short, the logarithmic score (LS) evaluates the forecast density, $p(y_t|\widetilde{\mathbf{y}})$, at the measured value, $\widetilde{y}_t$. This measure has a strong foundation in information theory (Shannon, 1948a,b; Kullback and Leibler, 1951) and is strictly local as it ignores model predicted probabilities of all non-realized outcomes. The continuous rank probability score (CRPS) is a quadratic measure of the difference between the predictive CDF, $F(y_t|\widetilde{\mathbf{y}})$, and the CDF, $G_t$, of $\widetilde{y}_t$. This is a common measure of performance for probabilistic forecasts of a scalar observation, $\widetilde{y}_t$, and rewards predictive distributions with a mass close to the observation. The pseudospherical score (PS) remunerates leptokurtic forecast densities as well, yet, this score rule acts on the predictive PDF and uses mixed powers in the numerator and denominator. For $\zeta = 2$ the PS reduces to the spherical score (SS) as used herein Friedman (1983). The interval score (IS) is important for quantile prediction, and rewards narrow prediction intervals. If the observation is outside the $[l_t, u_t]$ prediction interval, the score incurs a penalty, the size of which depends on the significance level, $\alpha$. Note that the IS evaluates only two predictive percentiles of the forecast distribution. This does not necessarily require an accurate description of the distribution of the predictand, $\widetilde{y}_t$. Hence, in the terminology of Gneiting and Raftery (2007), the IS is a *proper* but not a *strictly proper* scoring rule. The situation is different for the LS, CRPS and SS scoring rules, which are *strictly proper* and thus encourage forecasters to model correctly the entire predictive distribution. Thus, in principle, any of the LS, CRPS and SS rules will suffice to evaluate the quality of the probabilistic forecasts of the GL, GL$^+$, UL, SL and NL functions.

For each scoring rule, we compute a (time-averaged) mean score

$$\mathcal{S}(\mathbb{F}, \widetilde{\mathbf{y}}) = \frac{1}{n}\sum_{t=1}^{n}\mathcal{S}(F_t, \widetilde{y}_t), \tag{47}$$

where $\mathbb{F} = \{F_1, F_2, \ldots, F_n\}$ is the collection of empirical CDFs derived from the $m$ posterior realizations. Specifically, for $t \in (1, 2, \ldots, n)$, the empirical CDF (eCDF) of the predicted outcomes, $F_t$, is constructed from the $m$ entries, $(y^*_{t1}, y^*_{t2}, \ldots, y^*_{tm})$, in the $t^{\text{th}}$ row of the $n \times m$ matrix $\mathbf{Y}^*$ as follows

$$F_t(z) = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{z \geq y^*_{tj}\}, \tag{48}$$

where the indicator function, $\mathbb{1}\{a\}$, returns 1 if $a$ is true and zero otherwise. The eCDF is assumed continuous for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ and strictly positive on the bounded interval, $\Omega$. As a result, the density, $f_t(\cdot|\boldsymbol{\theta}, \boldsymbol{\delta})$

$$f_t(z) = \frac{\mathrm{d}}{\mathrm{d}z}[F_t(z)], \tag{49}$$

is continuous and strictly positive under the same support. The lower, $l_t$, and upper, $u_t$, limits of the $100(1-\alpha)\%$ prediction interval are equal to

$$l_t = F_t^{-1}(\alpha/2) \quad \text{and} \quad u_t = F_t^{-1}(1-\alpha/2), \tag{50}$$

where $F_t^{-1}(z)$ is the quantile function of $F_t(\cdot|\boldsymbol{\theta}, \boldsymbol{\delta})$.

The scoring rules of Table 5 are indispensable in our search for the *true* forecast densities, nevertheless, have not yet entered mainstream use in the hydrologic community. Therefore, we also consider other metrics of the forecast distribution hydrologists may be more familiar with (see Table 6). The reliability formalizes the thesis in probability theory that the sequence of probability integral transforms should consist of independent standard uniform random variables. In other words, if the $n$ observations are samples of the predictive distribution, $\mathbb{F}$, then the successive values of $F_t(\widetilde{y}_t)$

$$F_t(\widetilde{y}_t) = \int\!\!\int \mathbb{1}\{y^*_t(\boldsymbol{\theta}, \boldsymbol{\delta}) < \widetilde{y}_t\}\, p\bigl(y^*_t(\boldsymbol{\theta}, \boldsymbol{\delta})|\boldsymbol{\theta}, \boldsymbol{\delta}\bigr)\, p(\boldsymbol{\theta}, \boldsymbol{\delta}|\widetilde{\mathbf{y}})\mathrm{d}y^*_t(\boldsymbol{\theta}, \boldsymbol{\delta})\mathrm{d}(\boldsymbol{\theta}, \boldsymbol{\delta}), \tag{51}$$

should be uniformly distributed on the unit interval (Thyer et al., 2009; Renard et al., 2011). The reliability, RLBL, equals the mean absolute distance (1-norm) between the quantile function of the $F_t(\widetilde{y}_t)$'s and its counterpart of the standard uniform distribution. The multiplier of 2 scales the RLBL to the unit interval between 0 (poor) and 1 (perfect). The coefficient of variation, CV, measures the average sharpness of the predictive distribution. Smaller values of the CV are preferred. The coverage, $C$, equals the fraction of observations inside the prediction intervals. To be statistically meaningful and robust, $C$ should equal $1-\alpha$

[6] if $\mathcal{S}(F, \tilde{y}) > \mathcal{S}(G, \tilde{y})$ then $G$ is a better probabilistic forecast of the target variable, $\tilde{y}$, than $F$

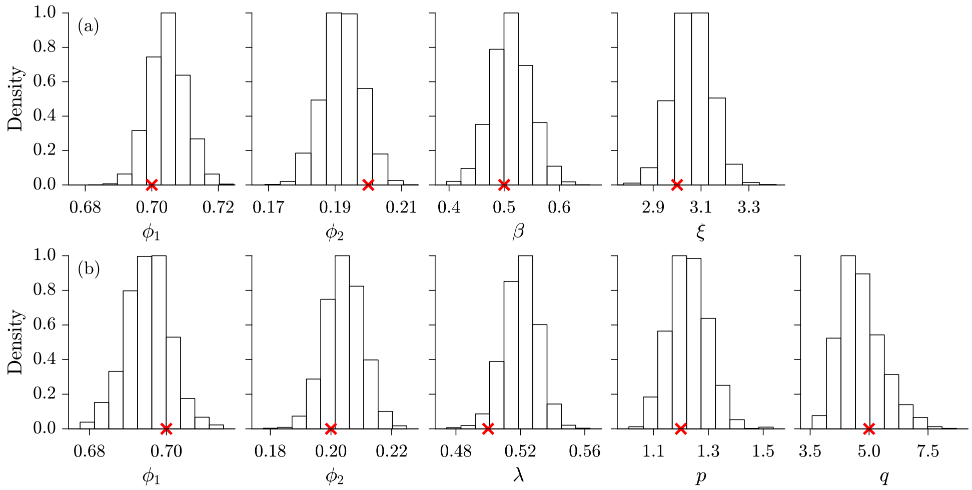


**Fig. 4.** Marginal distribution of the autoregressive coefficients, $\phi_1$ and $\phi_2$ of the training data record and coefficients (a) $\beta$ and $\xi$ of the SEP distribution and (b) $\lambda$, $p$ and $q$ of the SGT distribution. The red crosses portray the true values of the synthetic training data record.

at a significance level $\alpha$. The width, $W$, measures the average size of the $100(1-\alpha)$% prediction intervals.

The performance metrics of Table 6 measure different and complementary aspects of the forecast distribution. This complicates somewhat the ranking of the likelihood functions as we cannot combine the RLBL, CV, $W$ and $C$ into a single performance index without assigning arbitrary weights. One can instead sort the performance metrics according to the Pareto dominance principle and use the corresponding Pareto rank of each likelihood function as overall performance index (McInerney et al., 2017, 2019). But the width and coefficient of variation are properties of the predictive distribution only and, thus, do not guarantee honest forecasts. Furthermore, the reliability and coverage measure only two aspects of the statistical consistency between the distributional forecasts and the observations. In other words, these are necessary but not sufficient criteria for determining that the forecast distribution is accurate (see e.g. Hamill 2001). We therefore insist on using the *strictly proper* scoring rules of Table 5 for likelihood function evaluation and ranking.

### 5.5. General remarks

To simplify the discussion of our findings we classify the nuisance variables of the GL, GL$^+$, SL and UL functions in two different groups. Nuisance variables that are integral part of the underlying PDF of each likelihood function are coined internal nuisance variables. This includes the parameters of $f_{\text{SEP}}(a|\cdot)$, $f_{\text{SST}}(a|\cdot)$, and $f_{\text{SGT}}(a|\cdot)$ in Eqs. (9), (30) and (36), respectively. The remaining autoregressive coefficients, $\phi_1$ and/or $\phi_2$, will be referred to as external nuisance variables.

### 5.6. Case study I: An AR(2) process with non-Gaussian innovations

To benchmark the GL$^+$ and UL functions our first case study considers a simple autoregressive scheme

$$\widetilde{y}_t = \phi_1\widetilde{y}_{t-1} + \phi_2\widetilde{y}_{t-2} + \varepsilon_t \qquad (52)$$

with $\phi_1 = 0.7$, $\phi_2 = 0.2$ and $\varepsilon_t$'s drawn from (a) a standardized SEP distribution, $\varepsilon_t \sim \mathcal{SEP}(0,1,\beta,\xi)$, with $\beta = 0.5$ and $\xi = 3$ and (b) a standardized SGT distribution, $\varepsilon_t \sim \mathcal{SGT}(0,1,\lambda,p,q)$, with $\lambda = 0.5$, $p = 1.2$ and $q = 5$. We create a synthetic training data record, $\widetilde{\mathbf{y}} = [\widetilde{y}_1\ \widetilde{y}_2\ \dots\ \widetilde{y}_n]^\top$, with $n = 5{,}000$ using initial conditions, $\widetilde{y}_{-1} = 0$ and $\widetilde{y}_0 = 0$.

Fig. 4 presents histograms of the marginal posterior distributions of coefficients, $\phi_1$ and $\phi_2$, of the autoregressive model of Eq. (52) and the nuisance variables (a) $\beta$ and $\xi$ and (b) $\lambda$, $p$ and $q$ of the SEP and SGT distributions using the GL$^+$ and UL functions of Eqs. (26) and (41) with $s_0 = 1$. The "true" values of the coefficients are separately indicated with a red cross.

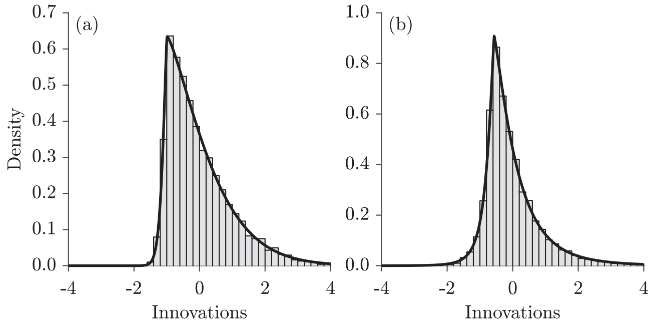


**Fig. 5.** Comparison of the true (black line) and MAP derived (gray bins) marginal distributions of the (a) SEP and (b) SGT innovations of Eq. (52).

The histograms of the SEP and SGT coefficients are well described by a normal distribution with mean approximately equal to the true values of the parameters of the AR(2) training data record and a small dispersion. This inspires confidence in the ability of the GL$^+$ and UL functions to correctly describe the marginal distribution of the partial studentized residuals. As the AR(2) operator of Eq. (12) does not preserve the skew and/or kurtosis of the SEP and SGT innovations, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, the mode of the marginal posterior distributions of the skew and/or kurtosis parameters do not necessarily have to coincide with their true values. But as we will show next, such small discrepancies do not have practical consequences.

Fig. 5 compares the empirical density function (histogram) of the partial residuals, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, of Eq. (52) with its theoretical (true) counterpart (solid line) for the (a) SEP and (b) SGT innovations, respectively.

Notice the excellent agreement between the theoretical PDF of the (a) SEP and (b) SGT innovations and their empirical frequency distribution derived from MCMC simulation using DREAM Suite (Vrugt, 2016). The different bins of the empirical histogram of the partial residuals provide an almost perfect characterization of the true distribution of the partial residuals. In fact, this holds for almost any other realization from the posterior parameter distribution. In other words, the posterior uncertainty in the empirical distribution of the innovations is very small.

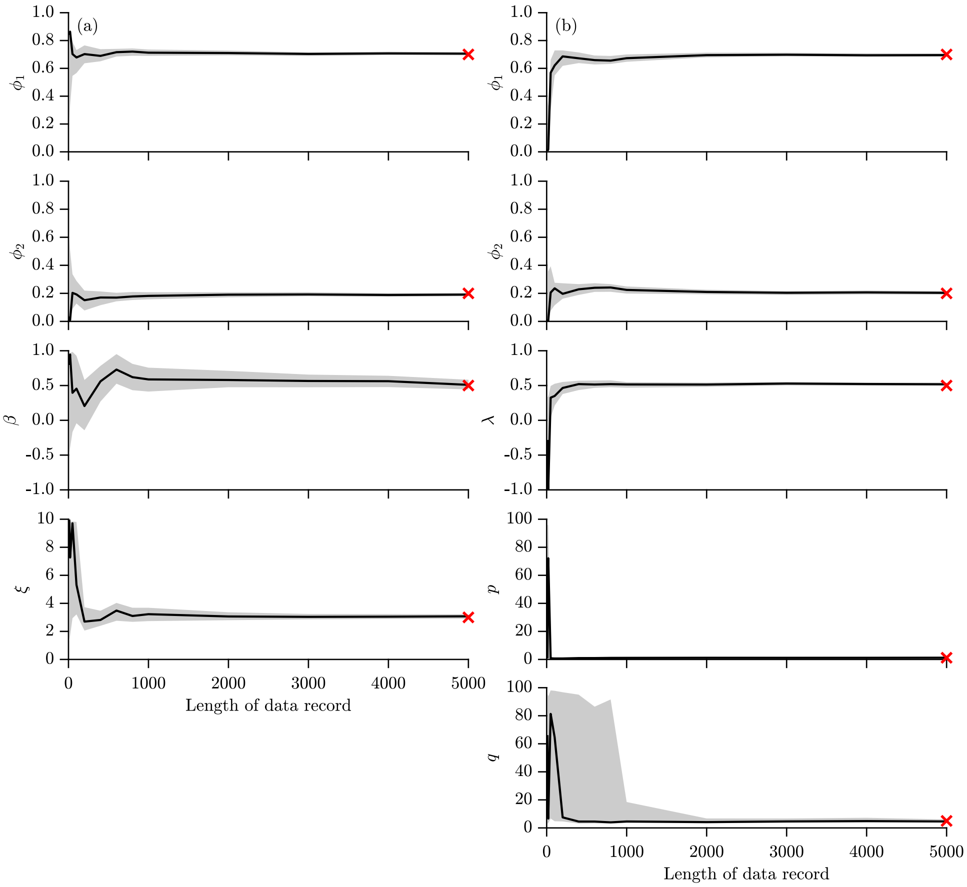


**Fig. 6.** Evolution of the MAP value (solid black lines) and its 95% confidence ranges (gray region) of the model parameters, $\phi_1$ and $\phi_2$, of the autoregressive scheme of Eq. (52) and the coefficients, (a) $\beta$ and $\xi$ of the GL$^+$ function, and (b) $\lambda$, $p$ and $q$ of the UL function. The true values of the parameters and nuisance variables are separately indicated with a red cross.

To understand the relationship between the length of the training data record and the posterior uncertainty of the coefficients of Eq. (52), please consider Fig. 6 which presents trace plots of the MAP values (solid black lines) of the two model parameters, $\phi_1$ and $\phi_2$, and nuisance variables of the (left) SEP and (right) SGT innovations. The gray region portrays the 95% confidence intervals of the MAP solution.

The autoregressive coefficients, $\phi_1$ and $\phi_2$, of the data-generating process of Eq. (52) converge rapidly to their theoretical values. After processing only one hundred data points, the 95% confidence intervals of the two autoregressive coefficients have collapsed to a small region immediately surrounding their MAP values. This conclusion holds for both the SEP and SGT innovations. Most of the nuisance variables of the GL$^+$ and UL functions converge at a somewhat smaller pace to their theoretical values with MAP values that go up and/or down the prior parameter space for small training data records. This is particularly true for the kurtosis parameters, $\beta$ and $q$. A training record of $n$ = 2,000 data points appears long enough to sufficiently constrain all the nuisance variables of the GL$^+$ and UL functions. Larger records hardly change the MAP values of the nuisance variables but further reduce the parameter uncertainty to a very small region interior to the uniform prior distribution. Intuitively, one would expect that the highest order moments of the partial residuals are most difficult to accurately characterize. Thus, it is no surprise that $\beta$, $q$ and/or $p$, may need a longer training data record, on average, than the skew parameters, $\xi$ and $\lambda$. Even a long training data record is no guarantee, however, that the skew and/or kurtosis parameters of the GL$^+$ and UL functions will converge exactly to their theoretical values. Such discrepancies do not harm the distribution of the partial residuals (see Fig. 5).

Altogether, the results of our first study demonstrate that the GL$^+$ and UL functions correctly infer the marginal distribution of the SEP and SGT innovations of the AR(2) scheme. This inspires confidence in the ability of the two likelihood functions to accurately describe the distribution of the partial residuals.

### *5.7. Case study II: Conceptual watershed model with measured discharge data*

As second case study, we illustrate the application of the GL, GL$^+$, SL and UL functions to a parsimonious 5-parameter conceptual watershed model. The HYdrologic MODel describes the rainfall–discharge relationship using five fictitious control volumes. These reservoirs simulate processes such as evaporation, percolation, river inflow and baseflow (see Fig. 7).

Hymod originates from the PhD thesis of Boyle (2001) and interested readers are referred to this publication for further details. We embed the process formulations of Fenicia et al. (2018) in a mass-conservative second-order integration method. Adaptive time stepping guarantee a robust and accurate numerical solution of the simulated fluxes and state variables. Table 7 presents the five hymod parameters with their corresponding symbols, units, and lower and upper bounds.

We illustrate the different likelihood functions by application to hydrologic data from the Leaf River near Collins, MS, USA. This medium-sized watershed with a strong winter regime according to the functional classification of Brunner et al. (2020) has been studied extensively in the hydrologic literature. We resort to the CAMELS data set (Newman et al., 2015; Addor et al., 2017) and simulate daily river discharge between 1 October 1998 and 30 September 2004 using daily estimates

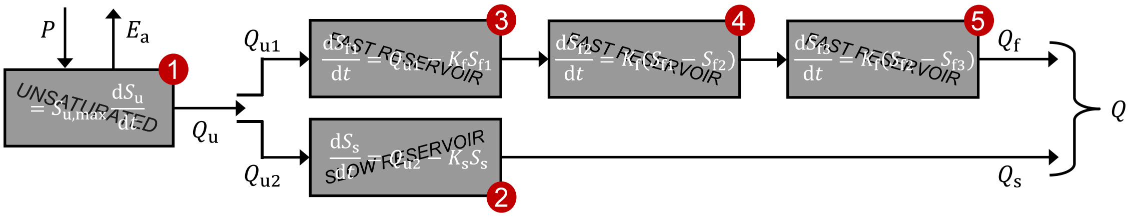

$P$
$E_a$
1
UNSATURATED
$= S_{u,max}\frac{dS_u}{dt}$
$Q_u$
$Q_{u1}$
$Q_{u2}$
3
FAST RESERVOIR
$\frac{dS_{f1}}{dt} = Q_{u1} - K_f S_{f1}$
4
FAST RESERVOIR
$\frac{dS_{f2}}{dt} = K_f(S_{f1} - S_{f2})$
5
FAST RESERVOIR
$\frac{dS_{f3}}{dt} = K_f(S_{f2} - S_{f3})$
2
SLOW RESERVOIR
$\frac{dS_s}{dt} = Q_{u2} - K_s S_s$
$Q_f$
$Q_s$
$Q$

**Fig. 7.** Schematic illustration of the HYdrologic MODel of Boyle (2001). Gray boxes, labeled in red, correspond to fictitious control volumes of the watershed which govern the rainfall–runoff transformation. The state variables, $S_u$, $S_s$, $S_{f1}$, $S_{f2}$ and $S_{f3}$, correspond to the water storage in each compartment. Arrows portray the fluxes into and out of the compartments, including precipitation, $P$, evaporation, $E_a$, precipitation converted into flow, $Q_u$, fast flow, $Q_f$, and baseflow, $Q_s$. The fluxes are computed as follows, $Q_u = P\left(1-(1-\overline{S}_u)^b\right)$, $E_a = E_p\overline{S}_u(1+c)/(\overline{S}_u+c)$, $Q_{u1} = aQ_u$, $Q_{u2} = (1-a)Q_u$, $Q_f = K_f S_{f3}$ and $Q_s = K_s S_s$, where $E_p$ signifies the potential evapotranspiration, $c = 10^{-2}$, $\overline{S}_u = S_u/S_{u,max}$ and $S_{u,max}$, $a$, $b$, $K_s$ and $K_f$ are unknown parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 7**
Summary of the hymod parameters and their symbols, units, and lower and upper bounds.

| Parameter | Symbol | Units | Min. | Max. |
|---|---|---|---|---|
| Maximum storage unsaturated zone | $S_{u,max}$ | mm | 50 | 1000 |
| Spatial variability of soil moisture capacity | $b$ | – | 0.1 | 10 |
| Flow partitioning coefficient | $a$ | – | 0 | 1 |
| Recession constant, slow reservoir | $K_s$ | 1/d | $10^{-5}$ | $10^{-1}$ |
| Recession constant, fast reservoir | $K_f$ | 1/d | $10^{-1}$ | 5 |

of catchment-averaged precipitation (Maurer product) and potential evapotranspiration derived from the formula of Oudin et al. (2005). Daily discharge measurements of the last five years on record (WY 1999–2004) are used for parameter estimation purposes using the GL, GL$^+$, SL and UL functions. This amounts to $n = 1827$ data points.

Table 8 lists the scoring rules and performance metrics of Tables 5 and 6 for the predictive streamflow distributions derived from the GL, GL$^+$, SL, UL and NL functions using different active sets of nuisance variables. Inactive nuisance variables are set to their default values (see Tables 1, 2 and 4). For completeness, the penultimate three columns list summary statistics of the goodness-of-fit of the hymod MAP simulation. Finally, the last column reports the Pareto rank of all likelihood functions but the GL function. These ranks are derived from non-dominated sorting of the *strictly proper* scoring rules of the tabulated formulations of the UL, GL$^+$, SL and NL functions. Thus, all likelihoods are ranked together treating LS, CRPS and SS as if they are complementary, non-commensurate, criteria. Rank one solutions are Pareto optimal; any other solution is considered inferior. A-priori one would advise using $\phi_1$ and/or $\phi_2$ in the active set of nuisance variables as the discharge residuals will almost surely exhibit serial correlation and, thus, fail an independence test. To explore to the fullest extent possible the performance of the different likelihood functions, we do include formulations without the external nuisance variables, $\phi_1$ and/or $\phi_2$. The row number in the first column serves as unique identifier of each likelihood function.

The tabulated data may be a bit overwhelming and, thus, we organize and discuss our results according to the goals and/or objectives of this paper.

*Comparison of GL and GL$^+$ functions.*

(i) The values of the scoring rules and performance metrics of the GL and GL$^+$ functions are in almost perfect agreement when the active set of nuisance variables is equivalent and made-up of only internal variables. Any small differences between the two likelihood functions are simply the result of a dissimilar treatment of the slope, $s_1$, of the measurement error function. The GL function considers $s_1$ a free parameter whose value is estimated jointly with the other nuisance variables. The GL$^+$ function, on the contrary, evokes the slope as a phantom variable so as to enforce a unit variance of the studentized residuals, $\underline{\mathbf{e}}(\theta,\delta)$. This latter approach frees $s_1$ from the list of nuisance variable in the GL$^+$, SL, NL and UL functions. This simplification is inconsequential as evidenced by the close match in the results of formulations 13, 15 and 16 of the GL$^+$ function and their equivalent counterparts 29, 31 and 32 of the GL function, respectively.

(ii) The scoring rules and performance metrics of the GL function deteriorate tremendously when the active set of nuisance variables includes one or more external variables, $\phi_1$ and/or $\phi_2$. This is an immediate consequence of the application of the AR($k$) process of Eq. (12) to non-homogenized residuals. This flaw was identified by Evin et al. (2013) and explains the rather inferior and overdispersed forecast distributions of the GL function as evidenced by the exceedingly high values of the scoring rules, relatively poor reliability, and unexpectedly large values of the coefficient of variation, width of the 95% prediction intervals and bias of the MAP simulation. Note that formulations 27 and 28 of the GL function even produce negative values of the CV. This exemplifies the deficient implementation of the GL function and demonstrates that we do not truncate negative discharge forecasts to zero. The GL$^+$ function does not suffer these problems as residual studentization precedes the treatment of serial correlation in Eq. (12). This results in a much better description of the discharge forecast distribution with prediction intervals that are much sharper on average and exhibit an improved coverage and reliability.

*Performance of distribution-adaptive likelihood functions.*

(i) The scoring rules and performance metrics vary considerably among the tabulated formulations of the same distribution-adaptive likelihood function. The choice of active nuisance variables exerts a large control on the performance of the UL, GL$^+$ and SL functions. The most advanced formulations do not necessarily yield the best overall performance. In other words, a larger number of active nuisance variables does not necessarily improve the performance of the UL, GL$^+$ and SL functions.

(ii) It is difficult, if not impossible, to single out a single best likelihood function and/or formulation thereof whose performance is uniformly excellent across all scoring rules and/or performance metrics. The GL$^+$ and SL functions achieve, on average, a somewhat lower Pareto rank than the UL function, suggesting that the advantages of this latter likelihood function did not fully materialize in the present example. Yet, we should not draw generalized conclusions from this single data set and model. We do not observe large differences in the performance of the GL$^+$, UL, SL and NL functions. None of their formulations consistently places best among the different scoring rules and/or performance metrics. Furthermore, care should be exercised that each non-dominated (= rank one) likelihood function satisfies residual assumptions. Such diagnostic checks of the residuals will impact substantially the ranking of the likelihood functions.

**Table 8**
Scoring rules and performance metrics of the predictive distributions of the UL, GL$^+$, SL, GL and NL functions for different active sets of nuisance variables. This is followed by summary statistics of the goodness-of-fit of the MAP solution and the Pareto rank of each likelihood function. Inactive nuisance variables assume default values.

| ID | Lik | Nuis. vars., $\boldsymbol{\delta}$ | Scoring rules, $\mathcal{S}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | | Performance metrics, $\mathcal{P}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | | Summary statistics | | | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LS | CRPS | SS | IS | RLBL | CV | $C$ | $W$ | $\mathcal{L}(\widehat{\theta},\widehat{\delta}\vert\widetilde{\mathbf{y}})$ | RMSE | PBIAS | |
| 1 | UL | $s_0,\lambda,p,q,\phi_1,\phi_2$ | 0.395 | 0.429 | −1.194 | 3.961 | 0.660 | 0.416 | 0.935 | 2.581 | 1507.0 | 1.573 | 26.91 | 5 |
| 2 | UL | $s_0,\lambda,\phi_1$ | 0.349 | 0.347 | −1.245 | 3.013 | 0.797 | 0.427 | 0.956 | 2.331 | 674.96 | 1.151 | 11.75 | 3 |
| 3 | UL | $s_0,p,\phi_1$ | 0.635 | 0.443 | −1.231 | 4.048 | 0.886 | 0.552 | 0.957 | 3.494 | 1418.1 | 1.645 | 27.54 | 6 |
| 4 | UL | $s_0,q,\phi_1$ | 0.397 | 0.444 | −1.301 | 3.644 | 0.858 | 0.482 | 0.950 | 2.797 | 1413.4 | 1.623 | 27.41 | 3 |
| 5 | UL | $s_0,\lambda,p,q$ | −0.055 | 0.334 | −1.446 | 3.761 | 0.976 | 0.594 | 0.948 | 2.483 | 107.68 | 1.141 | −4.120 | 1 |
| 6 | UL | $s_0,p,q,\phi_1$ | 0.168 | 0.440 | −1.273 | 3.809 | 0.881 | 0.513 | 0.961 | 3.197 | 1443.2 | 1.631 | 27.38 | 5 |
| 7 | UL | $s_0,p$ | 0.630 | 0.383 | −1.340 | 3.764 | 0.918 | 0.504 | 0.969 | 3.039 | −149.67 | 1.322 | 14.36 | 2 |
| 8 | UL | $s_0,\lambda,p,q,\phi_1$ | 0.272 | 0.429 | −1.199 | 3.967 | 0.664 | 0.415 | 0.935 | 2.584 | 1508.3 | 1.579 | 26.99 | 3 |
| 9 | UL | $s_0$ | 0.647 | 0.354 | −1.288 | 3.200 | 0.870 | 0.457 | 0.965 | 2.515 | −269.44 | 1.184 | 12.49 | 3 |
| 10 | GL$^+$ | $s_0,\beta,\xi,\phi_1,\phi_2$ | 0.809 | 0.449 | −1.305 | 3.796 | 0.918 | 0.506 | 0.970 | 3.166 | 1289.7 | 1.693 | 27.26 | 3 |
| 11 | GL$^+$ | $s_0,\beta,\phi_1$ | 1.007 | 0.453 | −1.297 | 3.870 | 0.922 | 0.517 | 0.970 | 3.241 | 1281.6 | 1.717 | 27.35 | 5 |
| 12 | GL$^+$ | $s_0,\xi,\phi_1$ | 0.404 | 0.345 | −1.271 | 3.248 | 0.891 | 0.475 | 0.958 | 2.532 | 663.09 | 1.143 | 9.298 | 3 |
| 13 | GL$^+$ | $s_0,\beta,\xi$ | −0.042 | 0.329 | −1.444 | 3.580 | 0.945 | 0.545 | 0.953 | 2.505 | 86.729 | 1.126 | −2.955 | 1 |
| 14 | GL$^+$ | $s_0,\beta,\xi,\phi_1$ | 0.702 | 0.448 | −1.301 | 3.799 | 0.917 | 0.506 | 0.970 | 3.170 | 1290.1 | 1.699 | 27.05 | 4 |
| 15[a] | GL$^+$ | $s_0,\beta$ | 0.656 | 0.383 | −1.341 | 3.769 | 0.919 | 0.505 | 0.970 | 3.043 | −149.51 | 1.320 | 14.05 | 2 |
| | | | 0.584 | 0.382 | −1.340 | 3.759 | 0.918 | 0.505 | 0.969 | 3.040 | −149.43 | 1.319 | 14.11 | |
| 16 | GL$^+$ | $s_0,\xi$ | 0.135 | 0.316 | −1.333 | 3.221 | 0.877 | 0.470 | 0.940 | 2.239 | −63.714 | 1.050 | 2.206 | 1 |
| 17 | GL$^+$ | $s_0$ | 0.620 | 0.355 | −1.288 | 3.200 | 0.869 | 0.457 | 0.964 | 2.513 | −269.28 | 1.170 | 12.26 | 3 |
| 18 | SL | $s_0,\nu,\xi,\phi_1,\phi_2$ | 0.128 | 0.438 | −1.295 | 3.840 | 0.789 | 0.447 | 0.935 | 2.559 | 1427.8 | 1.597 | 27.37 | 3 |
| 19 | SL | $s_0,\nu,\phi_1$ | 0.133 | 0.444 | −1.300 | 3.641 | 0.857 | 0.481 | 0.951 | 2.796 | 1413.5 | 1.632 | 27.33 | 3 |
| 20[a] | SL | $s_0,\xi,\phi_1$ | 0.364 | 0.348 | −1.244 | 3.011 | 0.796 | 0.427 | 0.957 | 2.330 | 674.90 | 1.127 | 12.05 | 4 |
| | | | 0.428 | 0.347 | −1.246 | 3.013 | 0.798 | 0.427 | 0.956 | 2.333 | 674.75 | 1.149 | 12.04 | |
| 21 | SL | $s_0,\nu,\xi$ | −0.057 | 0.333 | −1.431 | 3.748 | 0.969 | 0.599 | 0.941 | 2.362 | 104.23 | 1.143 | −4.525 | 1 |
| 22 | SL | $s_0,\nu,\xi,\phi_1$ | 0.129 | 0.438 | −1.288 | 3.869 | 0.791 | 0.455 | 0.938 | 2.562 | 1428.1 | 1.595 | 27.37 | 4 |
| 23 | SL | $s_0,\nu$ | 0.063 | 0.385 | −1.363 | 3.699 | 0.927 | 0.516 | 0.967 | 2.829 | −91.219 | 1.341 | 14.66 | 2 |
| 24 | SL | $s_0,\xi$ | 0.155 | 0.316 | −1.333 | 3.222 | 0.877 | 0.470 | 0.942 | 2.240 | −63.799 | 1.048 | 2.153 | 2 |
| 25 | SL | $s_0$ | 0.629 | 0.355 | −1.289 | 3.199 | 0.870 | 0.457 | 0.966 | 2.516 | −269.38 | 1.188 | 12.68 | 3 |
| 26 | GL | $s_0,s_1,\beta,\xi,\phi_1,\phi_2$ | 2.078 | 1.114 | −0.582 | 17.52 | 0.642 | 49.6 | 0.980 | 15.91 | 1642.4 | 1.856 | −60.56 | |
| 27[b] | GL | $s_0,s_1,\beta,\phi_1$ | 2.169 | 1.111 | −0.588 | 17.63 | 0.645 | −25.1 | 0.979 | 16.01 | 1642.3 | 1.806 | −58.82 | |
| 28[b] | GL | $s_0,s_1,\xi,\phi_1$ | 2.015 | 1.147 | −0.552 | 17.71 | 0.619 | −19.6 | 0.987 | 16.95 | 1207.2 | 1.550 | −42.16 | |
| 29[b] | GL | $s_0,s_1,\beta,\xi$ | −0.040 | 0.330 | −1.448 | 3.655 | 0.960 | 0.532 | 0.949 | 2.426 | 90.964 | 1.131 | −3.455 | |
| 30[b] | GL | $s_0,s_1,\phi_1$ | 2.024 | 1.161 | −0.546 | 17.95 | 0.616 | 72.1 | 0.987 | 17.15 | 1206.4 | 1.547 | −42.02 | |
| 31[b] | GL | $s_0,s_1,\beta,\xi,\phi_1$ | 2.087 | 1.110 | −0.584 | 17.38 | 0.641 | 353 | 0.980 | 15.76 | 1643.6 | 1.884 | −61.50 | |
| 32[b] | GL | $s_0,s_1,\beta$ | 0.930 | 0.384 | −1.344 | 3.694 | 0.922 | 0.482 | 0.968 | 2.896 | −147.10 | 1.332 | 14.65 | |
| 33[b] | GL | $s_0,s_1,\xi$ | 0.253 | 0.315 | −1.343 | 3.258 | 0.884 | 0.460 | 0.938 | 2.180 | −61.216 | 1.048 | 1.325 | |
| 34[a] | NL | $s_0,\phi_1$ | 0.399 | 0.436 | −1.082 | 3.336 | 0.700 | 0.437 | 0.957 | 2.627 | 639.00 | 1.348 | 22.31 | 6 |
| | | | 0.410 | 0.437 | −1.080 | 3.339 | 0.699 | 0.437 | 0.958 | 2.627 | 638.97 | 1.350 | 23.43 | |
| 35 | NL | $s_0$ | 0.653 | 0.355 | −1.288 | 3.197 | 0.870 | 0.458 | 0.964 | 2.515 | −269.46 | 1.188 | 12.46 | 4 |

[a] The two lines report the results of separate trials.

[b] Default values of $\beta = 0$, $\xi = 1$, $\phi_1 = 0$ and $\phi_2 = 0$ of Schoups and Vrugt (2010).

(iii) The SL function has the largest number of non-dominated formulations (two) and produces, on average, the lowest values of the IS and width, $W$, of the 95% prediction intervals. The GL$^+$ function yields the highest reliability. The UL function is a close competitor to the SL and GL$^+$ functions and achieves the highest mean value of the log-likelihood, $\mathcal{L}(\widehat{\theta},\widehat{\delta}|\widetilde{\mathbf{y}})$, but at the expense of a somewhat larger RMSE and PBIAS. The UL and GL$^+$ have only a single Pareto optimal formulation according to the *strictly proper* scoring rules of the discharge forecast distribution.

(iv) The UL, GL$^+$, SL functions produce equivalent scoring rules, performance metrics and summary statistics if the active set of nuisance variables includes only the intercept, $s_0$, of the measurement error function of Eq. (24). The performance of these three respective formulations (9, 17 and 25) is nearly indistinguishable and equal to that of the NL function with $s_0$ (= 35). This latter finding is enforced by the use of *normal* default values of the inactive nuisance variables and confirms our analytic derivations.

(v) The NL function with heteroscedastic measurement error function (35) achieves a performance that is at least comparable to that of most of the tabulated formulations of the GL$^+$, SL and UL functions. But as this formulation does not satisfy residual assumptions (shown later) we should benchmark the distribution-adaptive likelihood functions against the NL function with $s_0,\phi_1$ (= 34). All tabulated formulations of the UL, GL$^+$ and SL functions but the UL function with (3) $s_0,p,\phi_1$, improve upon the scoring rules and performance metrics of this normal likelihood as evidenced by their superior Pareto rank.

(vi) Different trials with the exact same likelihood function produce almost equal values of the scoring rules, performance metrics and summary statistics (see 15, 20 and 34). This demonstrates that the sampled chains successively converge to the approximately same target distribution. Thus, disparities in the tabulated statistics of the likelihood functions highlight differences between the UL, GL$^+$, SL and NL functions and/or their selection of active nuisance variables.

*On scoring rules and performance metrics.*

(i) The scoring rules of the forecast distribution display only weak relationships (see Tables F.1 and F.2 in Appendix F). The Pearson and Spearman correlation coefficients, $r$ and $\rho$, respectively, of LS, CRPS, SS and IS are consistently smaller than 0.5, except for $r_{\mathrm{CRPS,IS}} = 0.67$ and $\rho_{\mathrm{CRPS,IS}} = 0.62$, respectively. The weak correlation between the *strictly proper* scoring rules may seem counter-intuitive given that LS, CRPS and SS measure the divergence of the predictive distribution to the true forecast density. Apparently, this divergence can be expressed using three seemingly unrelated quantities. But this does not imply that the *strictly proper* scoring rules are independent. Their low degree of correlation simply articulates a large departure, on average, from the true forecast densities. We expect the pairwise correlations of LS, CRPS and SS to increase when the predictive distributions approximate more closely and consistently the true forecast densities. The relative independence of LS, CRPS and SS turns the selection of an adequate likelihood into a multi-criteria decision making problem.

(ii) The performance metrics of the forecast distribution do not display strong linear and monotonic relationships (Appendix F). The exception to this are the reliability and coefficient of variation, which have Pearson and Spearman correlation coefficients of 0.82 and 0.91, respectively. These two performance metrics measure different aspects of the forecast distribution, yet exhibit a relatively strong linear and rank dependence for the present data set and model.

(iii) Most of the scoring rules display only a weak relationship with the tabulated performance metrics. The strongest linear and rank correlation is found between the CRPS and width, $W$, of the forecast distribution ($r_{\text{CRPS},W} = 0.71$, $\rho_{\text{CRPS},W} = 0.87$) and the spherical score, SS, and the reliability ($r_{\text{SS,RLBL}} = -0.82$, $\rho_{\text{SS,RLBL}} = -0.76$). A somewhat weaker relationship is found between the SS and the coefficient of variation ($r_{\text{SS,CV}} = -0.73$, $\rho_{\text{SS,CV}} = -0.68$) and the logarithmic score, LS, and the coverage, $C$ ($r_{\text{LS},C} = 0.68$, $\rho_{\text{LS},C} = 0.73$). The scoring rules complement the performance metrics and help provide additional and/or orthogonal information about the quality of the forecast distribution.

(iv) All tabulated formulations of the UL, GL$^+$, SL and NL functions achieve the desired coverage of about 0.95 at a significance level, $\alpha = 0.05$. This performance metric does not discriminate among the different likelihood functions. Indeed, a coverage of $1-\alpha$ is simply a result of the rigorous application of probability theory. Note that a poor coverage alerts us to an erroneous implementation of the likelihood function, inaccurate characterization of the posterior $(\boldsymbol{\theta}, \boldsymbol{\delta})$-parameter distribution and/or use of an improper prior distribution.

(v) The CRPS demonstrates a rather small variation among the different likelihood formulations. This measure is commonly used when the probabilistic forecast is a CDF and the observations are scalars. This amounts to an eCDF, $G_t$, of each observation, $\widetilde{y}_t$, that is equal to a Dirac delta function with unit integral. The CRPS may show a larger variation among the different likelihoods if we replace this tall narrow spike with an actual CDF for each individual discharge observation. This CDF may be construed from replicates of discharge time series using the approach of Oliveira and Vrugt (2022). Alternatively, one can specify a normal CDF with variance derived from Eq. (24) using the MAP value of $s_0$. This latter approach is not ideal as the CDF of the data will depend on the likelihood function used.

(vi) The reliability, coefficient of variation, width and coverage measure specific properties of the forecast distribution deemed important to represent accurately in practical application. These performance metrics have the advantage of being easy to analyze and interpret, but as necessary and insufficient conditions of the predictive distribution these metrics do not help in the pursuit of the true forecast density. The same is true for *proper* scoring rules, such as the interval score.

(vii) The *strictly proper* scoring rules provide a sufficient description of the true forecast density, yet, the LS, CRPS and SS are difficult to interpret as they provide an integrated measure of the divergence of the predictive distribution from the true forecast density.

*The treatment of residual autocorrelation.*

(i) The use of external nuisance variables, $\phi_1$ and/or $\phi_2$, in the UL, GL$^+$, SL and NL functions, enhances substantially the log-likelihood of the posterior realizations but at the expense of a deterioration in most (if not all) of the scoring rules and performance metrics. Specifically, the treatment of serial correlation reduces the reliability of the discharge forecast distribution and increases the width, $W$, of its 95% prediction intervals. In interpreting this finding it is important to reiterate that the autoregressive operator of Eq. (12) is evaluated in simulation mode using the recipe detailed in Algorithm E.1. If we switch to one-observation ahead forecasting and submit the actual residuals to the autoregressive operator instead, then the quality of the predictive discharge distribution would improve considerably, beyond the scoring rules and performance metrics obtained with internal nuisance variables only. However, this would obfuscate the comparison of AR and non-AR based formulations of the likelihood function(s) and, thus, we do not present these results herein.

(ii) The use of external nuisance variables, $\phi_1$ and/or $\phi_2$, substantially increases the RMSE and percentage bias of the MAP simulation. This is a result of model evaluation in simulation mode. The MAP discharge simulation equals the deterministic output of hymod only, thus, does not include the AR-extension of Eq. (12).

(iii) There are no benefits to using an AR(2) model of the studentized raw residuals. The value of $\phi_2$ consistently converges to zero in all respective formulations. Higher-order autoregressive models may be required for other data sets and/or models to remove serial correlation among the studentized raw residuals.

Altogether, the results confirm that the GL$^+$ function is superior to the GL function and has the additional benefit of requiring the estimation of one less nuisance variable. The use of studentized residuals in the autoregressive model of Eq. (12) rectifies a critical deficiency of the GL function and guarantees an accurate implementation of the SEP log-likelihood of Eq. (26). This makes obsolete the GL function and, thus, we focus our attention on the UL, GL$^+$ and SL functions.

The *strictly proper* scoring rules facilitate a rigorous evaluation of the forecast distributions derived from the posterior realizations of the different likelihood functions. This reduces the need for a visual assessment of the results. The LS, CRPS and SS (and/or IS) should not be used as sole determinant for likelihood function selection as these scoring rules do not verify whether residual assumptions are satisfied. Some of the likelihood functions perform better than others, but no formulation consistently ranks best among the different scoring rules and/or performance metrics. To illustrate this in more detail, please consider Fig. 8 which presents bivariate scatter plots of the values of the *strictly proper* scoring rules of the UL, GL$^+$, SL and NL functions of Table 8. We assign a separate color to each likelihood function and use empty (white) circles to differentiate formulations that do not treat serial correlation from those that do (filled circles).

The scatter diagrams are only cross-sections of the three-dimensional space of *strictly proper* scoring rules, yet highlight some of the difficulties involved in the selection of a suitable likelihood function. The dotty plots confirm our earlier claim that the treatment of serial correlation tends to deteriorate, on average, the quality of the discharge forecast distribution. This testifies to the use of simulated residuals in the autoregressive operator of Eq. (12). The different graphs do not favor a single unique formulation of the likelihood function, but rather demonstrate the presence of multiple different optimal formulations. These so-called Pareto solutions represent trade-offs among the different scoring rules and have the property that moving from one solution to another results in the improvement of one scoring rule while causing a simultaneous deterioration in one or more others. This is a result of the low correlation among the *strictly proper* scoring

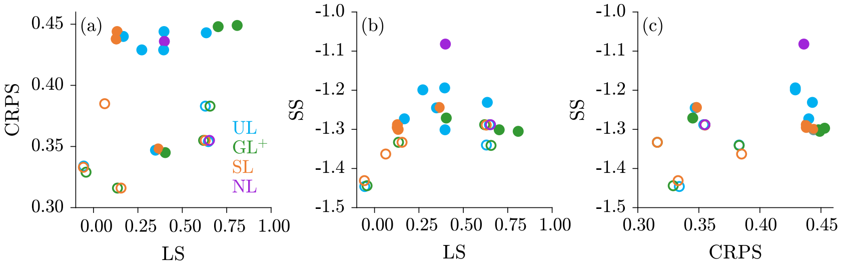


**Fig. 8.** Two-dimensional scatter plots of the values of the scoring rules derived from the tabulated formulations of the GL$^+$, SL, UL and NL functions: (a) LS – CRPS, (b) LS – SS, and (c) CRPS – SS. The color of each data point matches to the likelihood function. The empty circles correspond to the formulations of Table 8 that do not include external nuisance variables, $\phi_1$ and/or $\phi_2$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

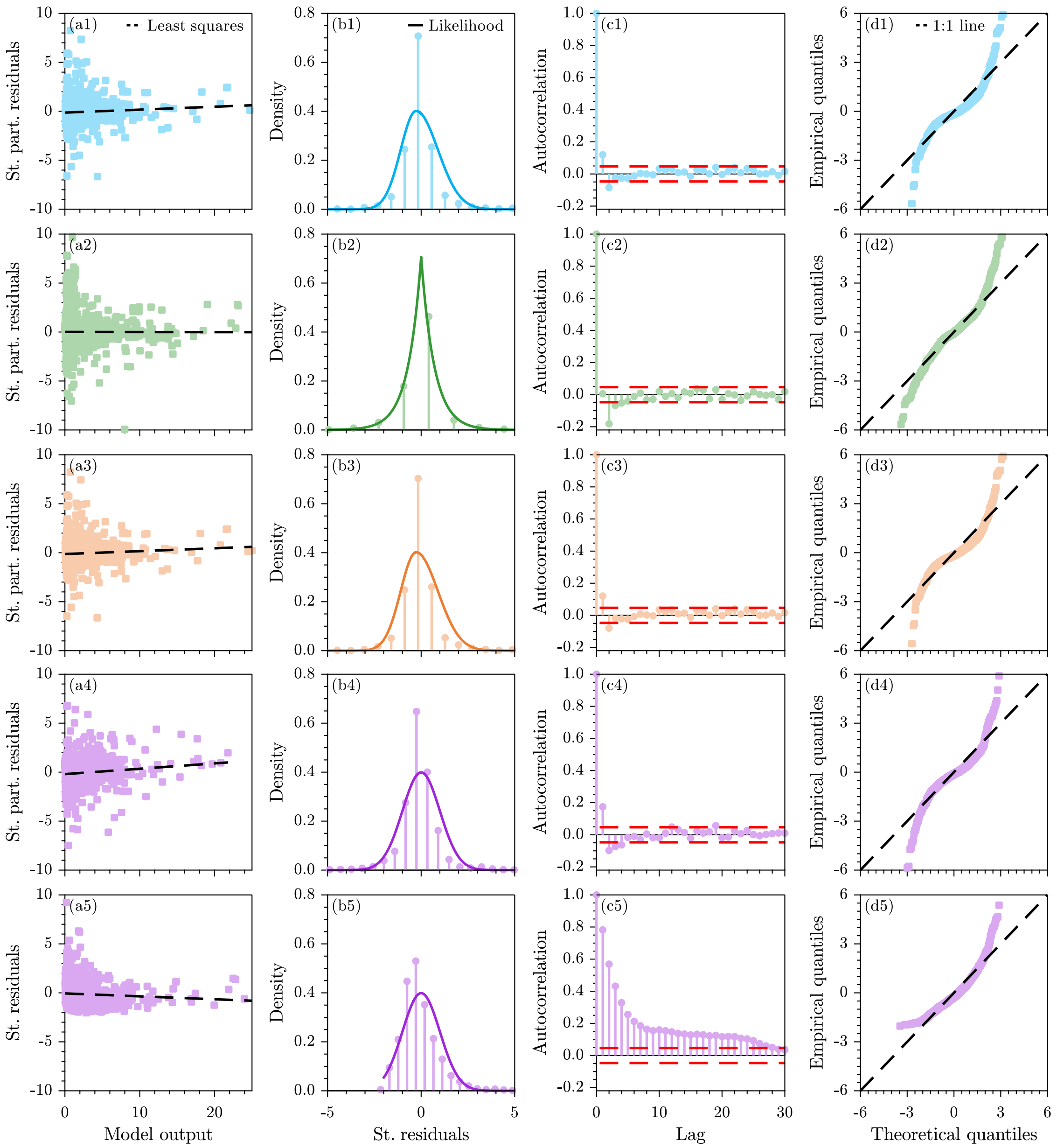


**Fig. 9.** Diagnostic checks of the residuals: (a) scatter plot, (b) histogram, (c) sample autocorrelation function and (d) quantile–quantile plot of the standardized partial residuals, $\underline{\epsilon}(\theta, \delta)$, of the hymod MAP simulation for our selection of likelihood functions, including (1) the UL function with $s_0, \lambda, \phi_1$, (2) the GL$^+$ function with $s_0, \xi, \phi_1$, (3) the SL function with $s_0, \xi, \phi_1$, and the NL function with (4) $s_0, \phi_1$, and (5) $s_0$, respectively. The dashed lines portray (a) the least squares fit, (c) the 95% confidence limits of white noise, the so-called (Anderson, 1941) limits, and (d) 1:1 line of the theoretical and empirical quantiles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

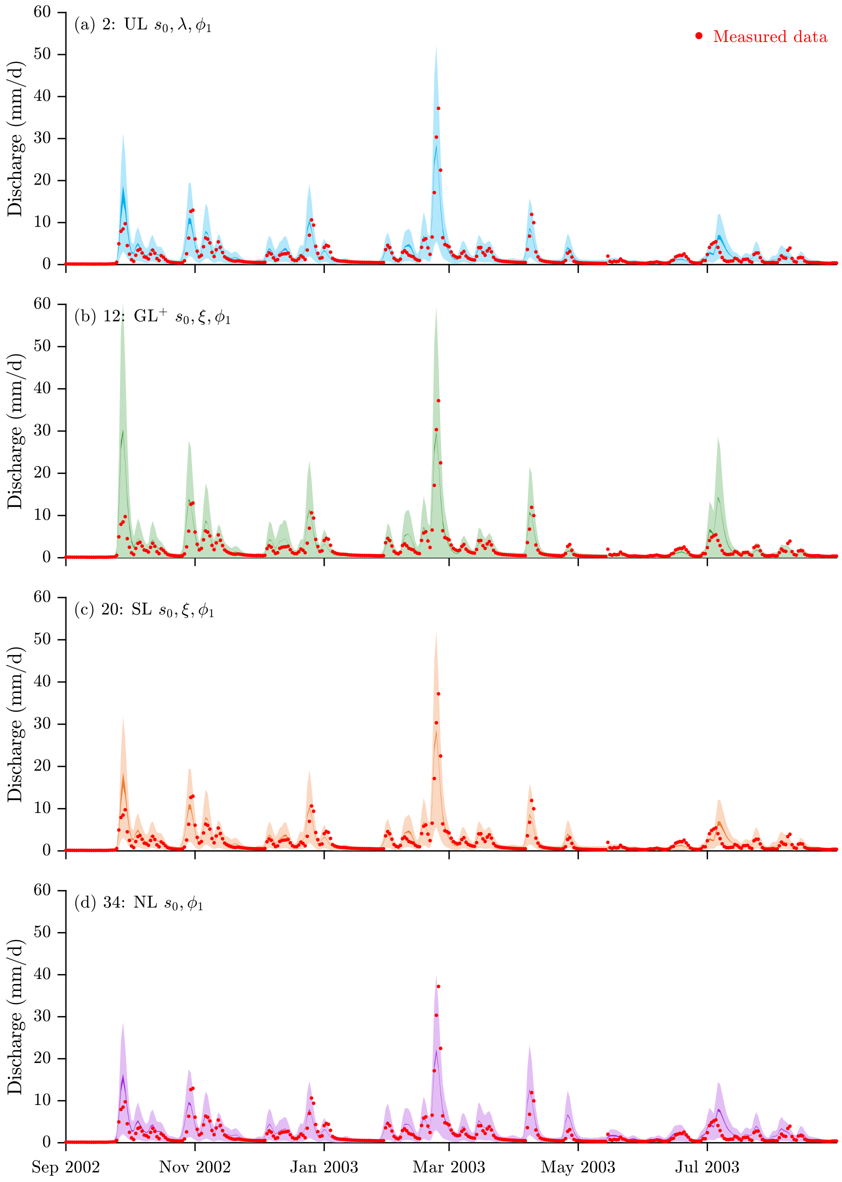


**Fig. 10.** Time series plot of observed (red dots) and simulated discharge: (a) UL: $s_0, \lambda, \phi_1$, (b) GL$^+$: $s_0, \xi, \phi_1$, (c) SL: $s_0, \xi, \phi_1$, and (d) NL: $s_0, \phi_1$ for a representative 365-day period for the Leaf river basin. The dark and light shaded regions portray the 95% discharge uncertainty intervals due to parameter and total uncertainty, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

rules, and consequently, we can treat the selection of an adequate likelihood function as a multi-criteria decision-making problem. The spatial arrangement of the different colors in each dotty plot confirm our earlier finding that the different formulations of the SL function achieve, on average, the best performance, followed by the GL$^+$ and UL functions. Upon inspection of the Pareto ranks of the AR-based formulations of the GL$^+$, UL and SL functions listed in Table 8, we conclude that the UL function with (1) $s_0, \lambda, p, q, \phi_1, \phi_2$, (2) $s_0, \lambda, \phi_1$, (4) $s_0, q, \phi_1$, (6) $s_0, p, q, \phi_1$ and (8) $s_0, \lambda, p, q, \phi_1$, the GL$^+$ function with (10) $s_0, \beta, \xi, \phi_1, \phi_2$, (11) $s_0, \beta, \phi_1$, (12) $s_0, \xi, \phi_1$, (14) $s_0, \beta, \xi, \phi_1$, and the SL function with (18) $s_0, \nu, \xi, \phi_1, \phi_2$, (19) $s_0, \nu, \phi_1$, (20) $s_0, \xi, \phi_1$ and (22) $s_0, \nu, \xi, \phi_1$ achieve a performance that is superior to the NL function of Eq. (27) with $s_0, \phi_1$ (purple dot). At this time it is important to reiterate that we could have also ranked the different likelihood functions according to any of the three *strictly proper* scoring rules. Then likelihood functions 18, 12 and 10, would have come on top for the LS, CRPS and SS rules, respectively, and the so-ordered likelihood functions would significantly overlap with the Pareto dominance-based selection listed above. Of these solutions, we will investigate in more detail next, the UL function with (2) $s_0, \lambda, \phi_1$, the GL$^+$ function with (12) $s_0, \beta, \phi_1$ and SL function with (20) $s_0, \xi, \phi_1$.

The scoring rules and/or performance metrics do not tell us whether residual assumptions are satisfied and whether hymod accurately describes the measured discharge record. Next, we verify whether the partial residuals of the UL, GL$^+$, SL and NL functions satisfy assumptions of homogeneity and independence (among others). Fig. 9 presents the results of our analysis. The left graph displays the studentized partial residuals as a function of the hymod simulated discharge. The second graph shows the assumed (solid line) and actual (vertical lines) PDF of the studentized partial residuals. The third graph presents the sample autocorrelation function of the studentized partial residuals. The fourth and last graph compares theoretical and empirical quantiles of the studentized partial residuals. To benchmark our findings, the bottom panel visualizes the results of the NL function with $s_0$ (= 35).

This likelihood function does not treat serial correlation, thus, we expect that it will violate residual assumptions.

The scatter plots appear to suggest that the studentized partial residuals are not homogeneous, but rather decrease in magnitude with increasing value of the simulated flows. But this is an optical illusion. Upon close inspection of the data scattering at different flow levels, we can reveal that the variance of the studentized (partial) residuals is rather constant and independent of the magnitude of the simulated discharge. The near-zero slopes of the regression lines confirm the homogeneous nature of the $\underline{\varepsilon}_f(\theta, \delta)$'s. The empirical density functions of the partial residuals follow a Laplacian distribution that is generally well described by the respective PDFs of the likelihood functions. The characteristic peak of the partial residuals is only duplicated by the $\mathrm{GL}^+$ function. Note that the PDFs of the UL and SL functions exhibit a small skew to the right, whereas the PDF of the $\mathrm{GL}^+$ function is symmetric around zero. By definition, the PDFs of the NL function are without skew. Autocorrelation is negligible at all lags for likelihood functions 2, 12, 20 and 35. This is an immediate consequence of the use of the AR operator in Eq. (12). The NL function in the bottom panel does not treat serial correlation, hence, the studentized residuals violate the independence assumption. The theoretical and empirical quantiles of the studentized partial residuals are in close agreement within the interquartile range. Outside this interval the empirical quantiles of the likelihood functions deviate considerably from the 1:1 line. Indeed, none of the likelihood functions characterize well the tails of the residual distribution. This involves only a small number of residuals, but these residuals are critically important in describing adequately the prediction limits of the simulated discharge.

Fig. 10 presents time series plots of the observed (red dots) and simulated discharge of the $m$ posterior realizations for a representative 1-year period of the training record of the Leaf river basin using likelihood functions (a) 2, (b) 12, (c) 20, and (d) 34. The dark and light shaded regions summarize the 95% credible intervals due to parameter and total uncertainty, respectively.

For all four likelihood functions the hymod parameter uncertainty makes up an insignificant part of the prediction uncertainty with 95% credible intervals that collapse to a (dark) line with the exception of a few peak flows. The discharge prediction uncertainty is largest for the $\mathrm{GL}^+$ function with 95% intervals that extent to zero for the largest streamflows. This is visually not particularly pleasing. The NL function with $s_0, \phi_1$ provides a relatively good description of the uncertainty with MAP solution (not shown) that tracks reasonably well the measured discharge record. The prediction uncertainty of the UL and SL functions depend most on the simulated flow level. The discharge uncertainty is relatively small at low flows but increases substantially during rainfall events with increasing discharge. The prediction uncertainty of the NL function appears more constant with a comparatively large uncertainty at low flows and comparatively small uncertainty at the peak flows. Altogether, the prediction uncertainty of the UL function is visually most pleasing with 95% intervals that are smallest on average.

Finally, we investigate the effect of the choice of likelihood function on the posterior marginal distribution of the hymod parameters (see Fig. 11).

The hymod parameters appear well defined by the measured discharge. With exception of the NL function with $s_0, \phi_1$ we observe a close agreement in the marginal parameter distributions derived from the different likelihood functions. This is particularly true for the UL and SL functions. We see nicely bell-shaped histograms with an approximately similar mean and comparably small spread. Variations in the mean of each hymod parameter among the different likelihood functions are relatively small, certainly compared to the prior ranges of the parameters. The hymod parameters exhibit a negligible correlation, with exception of $S_{\mathrm{u,max}}$ and $b$ of the NL function with $s_0, \phi_1$. This explains their much enlarged uncertainty. These results confirm that the choice of likelihood function has an effect on the posterior parameter values.

Note, in our analysis here we limit our attention to the performance of hymod during the 5-year calibration period and do not present performance statistics and/or analyze results for an independent evaluation period. The calibration period suffices to illustrate the application of the different likelihood functions and the comparison and/or ranking of their predictive distributions. Furthermore, for a long data set and parsimonious model such as hymod, we do not expect large changes to the performance of the likelihood functions in an independent evaluation period under the stationarity thesis. If model selection is of interest then the user can resort to GAME sampling (Volpi et al., 2017). This method numerically integrates the posterior distribution in pursuit of the so-called marginal likelihood. This measure suffices to determine which model and/or likelihood is most supported by the experimental (discharge) data.

## 6. Conclusions

This paper was concerned with the formulation of herein called distribution-adaptive likelihood functions in the application of Bayesian epistemology to uncertainty quantification of hydrologic watershed models. This class of likelihood functions does not require prior assumptions about the expected distribution of the residuals. Inference takes place over the model parameters and space of distribution functions defined by the nuisance variables of the likelihood function.

The goals of this paper were threefold. First, we presented a revised formulation of the generalized likelihood (GL) function of Schoups and Vrugt (2010). By enacting the treatment of serial correlation on the studentized raw residuals rather than the raw residuals, this so-called $\mathrm{GL}^+$ function rectifies a critical deficiency of the GL function. This correction guarantees an accurate computation of the SEP log-likelihood and allows for a much more robust joint inference of the autoregressive coefficients and variance, skew and/or kurtosis of the residuals. As secondary goal, we presented a further generalization of the $\mathrm{GL}^+$ function, coined the universal likelihood (UL) function. This likelihood function uses as its main building block the skewed generalized Student's $t$-distribution of Theodossiou (2015) and extends applicability to a much larger family of well-known probability distributions. As third and last goal we introduced the use of *strictly proper* scoring rules to evaluate, compare and rank the predictive distributions of different likelihood functions. The logarithmic, continuous rank probability and spherical score rules condense the accuracy of a distribution forecast to a single value while retaining attractive statistical properties and incentivizing a correct modeling of the predictive distribution.

The power and usefulness of the $\mathrm{GL}^+$ and UL functions were demonstrated by application to two case studies. The first study involved a simple autoregressive data generating process with SEP and/or SGT innovations and was used to investigate the convergence properties of the $\mathrm{GL}^+$ and UL functions. Both likelihood functions converge upon the true values of the model parameters and shape coefficients of the distribution of the innovations. The uncertainty in the autoregressive coefficients and nuisance variables of the partial residuals dissipates quickly with increasing length of the training data record. As the AR(2)-operator does not preserve higher-order moments of the SEP and/or SGT innovations, the mode of the marginal distributions of the skew and/or kurtosis parameters may not coincide exactly with their true values. Such discrepancies, however, are inconsequential as the distribution of the innovations is perfectly described.

The second case study considered the application of distribution-adaptive likelihood functions to a simple five-parameter watershed model named hymod using daily discharge measurements of the Leaf

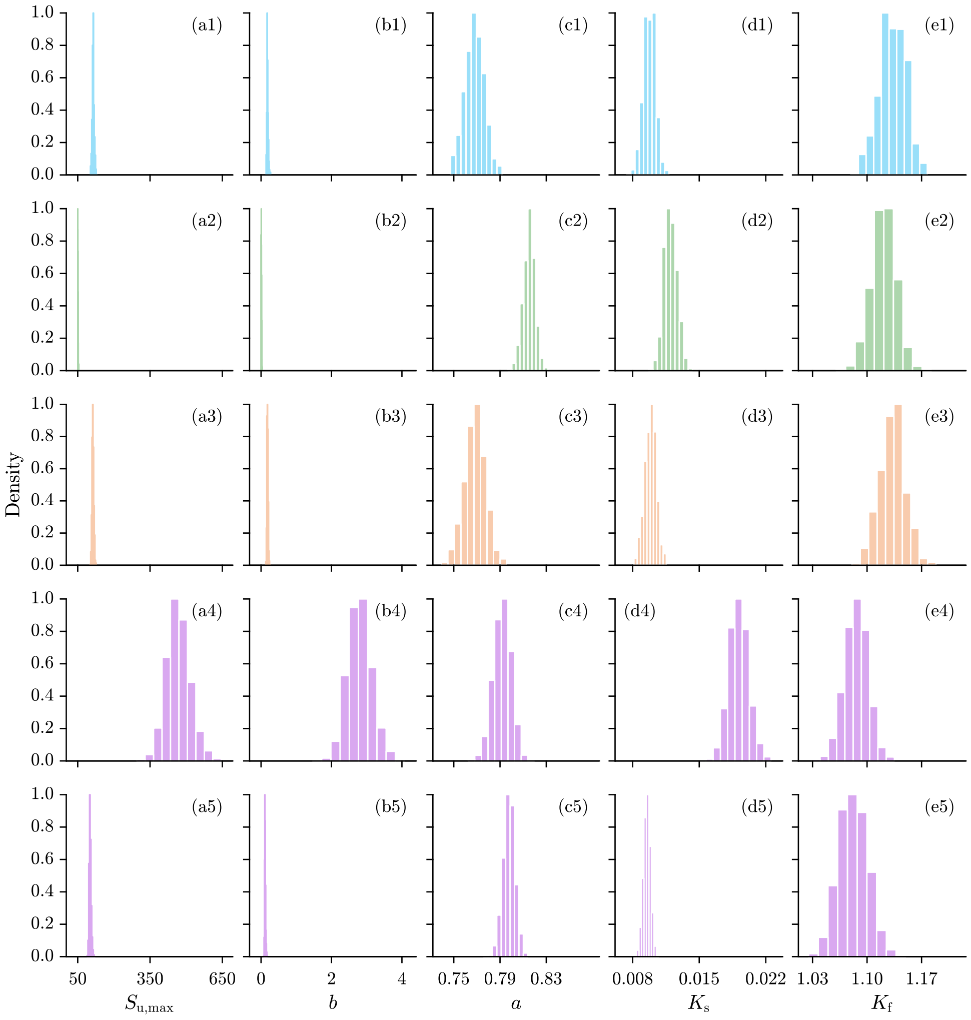


**Fig. 11.** Histograms of the marginal posterior distributions of the hymod parameters, (a) $S_{\mathrm{u,max}}$, (b) $b$, (c) $a$, (d) $K_{\mathrm{s}}$ and (e) $K_{\mathrm{f}}$ for likelihood functions, (1) 2, (2) 12, (3) 20, (4) 34 and (5) 35. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

river. Different formulations of the GL$^+$ and UL functions were evaluated using the logarithmic, continuous rank probability and spherical scoring rules and benchmarked against the GL function, the Student $t$-likelihood (SL) of Scharnagl et al. (2015) and a normal likelihood (NL) with first-order autoregressive model and/or heteroscedastic measurement errors. Our most important results were as follows.

1. The GL$^+$ function is superior to the GL function and has the advantage of requiring the specification of one less nuisance variable.
2. The performance of the GL$^+$, SL and UL functions depends in large part on the choice of nuisance variables. The most complex formulations of the GL$^+$, SL and UL functions (all nuisance variables in active set) do not yield the best performance.
3. The internal nuisance variables of the GL$^+$, SL and UL functions enable a better description of the discharge residuals. This reduces the width of the discharge prediction intervals.
4. The normal likelihood achieves a performance that is fairly comparable to several formulations of the GL$^+$, SL and UL functions.
5. The treatment of autocorrelation deteriorates the values of the scoring rules and performance metrics of the forecast distribution. If, however, we switch to one-observation ahead forecasting and work with actual rather than simulated residuals, then the use of autocorrelation improves substantially the quality of the predictive discharge distribution beyond the scoring rules and performance metrics obtained for internal nuisance variables only (e.g. see Evin et al., 2014).
5. Scoring rules are indispensable in our search for the true forecast distribution
6. *Strictly proper* scoring rules facilitate a statistically rigorous evaluation and ranking of different likelihood functions.
7. We cannot single out a single best likelihood function whose performance is uniformly excellent across all scoring rules and performance metrics of the forecast distribution.
8. The use of one *strictly proper* scoring rule suffices, in principle, for evaluation and ranking of the likelihood functions. The simultaneous use of the logarithmic, continuous rank probability and spherical scores converts the selection of an adequate likelihood function into a multi-criteria problem. Care should be exercised that each likelihood function satisfies residual assumptions.
9. The Pareto optimal likelihood functions produce fairly similar marginal distributions of the hymod parameters.

Altogether, our results for hymod favored the use of a distribution-adaptive likelihood function in describing the rainfall–discharge dynamics of the Leaf River. A leptokurtic distribution of the partial residuals provided the overall best characterization of the measured discharge record and associated uncertainties.

## CRediT authorship contribution statement

**Jasper A. Vrugt:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Debora Y. de Oliveira:** Validation, Methodology, Formal analysis, Resources, Visualization. **Gerrit Schoups:** Methodology, Validation, Investigation. **Cees G.H. Diks:** Methodology, Validation.

## 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.

## Data availability

Data will be made available on request.

## Acknowledgment

The comments of Carlos Lima and one anonymous reviewer led to a significantly improved manuscript. The second author gratefully acknowledges the financial support received from the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), grant number 88881.174456/2018-01.

## Appendix A. Standardized skewed exponential power (SEP) density

We revisit Appendix A of Schoups and Vrugt (2010) and reiterate how the standardized skew exponential power density function in Eq. (9) can be obtained from the exponential power (EP) or generalized normal distribution of Subbotin (1923) and Box and Tiao (1992) using the method of Fernandez and Steel (1998).

The density of the standardized EP distribution (zero-mean and unit variance) at point $a$ may be computed as follows (Box and Tiao, 1992)

$$f_{\mathrm{EP}}(a|0,1,\beta) = \omega_\beta \exp\bigl(-c_\beta |a|^{2/(1+\beta)}\bigr), \tag{A.1}$$

where $\beta \in (-1,1]$ is a so-called kurtosis parameter which controls the peakedness of the distribution[7] and $|\cdot|$ denotes the absolute value or modulus operator. The values of the scalars $c_\beta$ and $\omega_\beta$ depend on the kurtosis parameter and are given by Box and Tiao (1992)

$$c_\beta = \left(\frac{\Gamma\bigl(3(1+\beta)/2\bigr)}{\Gamma\bigl((1+\beta)/2\bigr)}\right)^{1/(1+\beta)} \tag{A.2a}$$

$$\omega_\beta = \frac{\Gamma^{1/2}\bigl(3(1+\beta)/2\bigr)}{(1+\beta)\Gamma^{3/2}\bigl((1+\beta)/2\bigr)}, \tag{A.2b}$$

where $\Gamma(b)$ signifies the incomplete Gamma function evaluated at $b$

$$\Gamma(b) = \int_0^\infty x^{b-1}\exp(-x)\mathrm{d}x \qquad \forall b \in \mathbb{R}_+, \tag{A.3}$$

which satisfies the recursion $\Gamma(b+1) = b\Gamma(b)$. The gamma function can be approximated numerically

$$\log\bigl(\Gamma(b)\bigr) = -\gamma b - \log(b) + \sum_{m=1}^{\infty}\left(\frac{b}{m} - \log\left(1+\frac{b}{m}\right)\right), \tag{A.4}$$

where $\gamma = \lim_{m\to\infty}\left(-\log(m) + \sum_{k=1}^{m}\frac{1}{k}\right) \approx 0.5772$ is the so-called Euler–Mascheroni constant and $\log(\cdot)$ signifies the natural logarithm.

The EP density, $f(a|\beta)$, in Eq. (A.1) is symmetric around $a = 0$. This symmetry impairs our ability to describe accurately skewed residual distributions with an upper or lower tail. Fernandez and Steel (1998) developed a general solution for the treatment of skew in symmetric distributions with closed-form mathematical expressions. Their template density, $f_{\mathrm{skew}}(a|\xi)$, with skewness parameter, $\xi \in \mathbb{R}_+$, introduces tails in an arbitrary symmetric density, $f(\cdot)$, as follows

$$f_{\mathrm{skew}}(a|\xi) = \frac{2}{\xi+\xi^{-1}} f\bigl(a\xi^{-\operatorname{sign}(a)}\bigr) = \frac{2}{\xi+\xi^{-1}} f\left(\frac{a}{\xi^{\operatorname{sign}(a)}}\right), \tag{A.5}$$

where $\operatorname{sign}(\cdot)$ is the signum function

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases} \tag{A.6}$$

and the mean, $\mu_\xi$, and variance, $\sigma_\xi^2$, of $a$ satisfy the following relationship (Fernandez and Steel, 1998)

$$\mu_\xi = M_1(\xi - \xi^{-1}) \tag{A.7a}$$

$$\sigma_\xi^2 = (M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2, \tag{A.7b}$$

wherein $M_j$ is the $j^{\mathrm{th}}$ absolute moment of the standardized EP density, $f_{\mathrm{EP}}(\cdot|0,1,\beta)$, of Eq. (A.1)

$$M_j = \mathbb{E}(|x^j|) = \int_{-\infty}^{\infty} x^j f_{\mathrm{EP}}(x|0,1,\beta)\,\mathrm{d}x. \tag{A.8}$$

We can now substitute the exponential power density, $f_{\mathrm{EP}}(a|0,1,\beta)$, of Eq. (A.1) into the skew distribution, $f_{\mathrm{skew}}(a|\xi)$, of Eq. (A.5). The result is a non-standardized skewed exponential power (SEP) density with mean, $\mu_\xi$, and variance, $\sigma_\xi^2$, of $a$ determined by Eqs. (A.7a) and (A.7b). To negate changes in the mean and variance of $a$ imposed by the use of the skew density of Eq. (A.5), we must scale the SEP density, and, therefore, use as its input argument, $\mu_\xi + \sigma_\xi a$, rather than $a$. This leads to the following expression for the standardized SEP density (Schoups and Vrugt, 2010)

$$f_{\mathrm{SEP}}(a|0,1,\beta,\xi) = \frac{2\sigma_\xi\omega_\beta}{\xi+\xi^{-1}} \exp\left(-c_\beta\left|\frac{\mu_\xi+\sigma_\xi a}{\xi^{\operatorname{sign}(\mu_\xi+\sigma_\xi a)}}\right|^{2/(1+\beta)}\right), \tag{A.9}$$

where $c_\beta$, $\omega_\beta$, $\mu_\xi$, and $\sigma_\xi$, are a function of the kurtosis, $\beta$, and skewness, $\xi$, using Eqs. (A.2a), (A.2b), (A.7a) and (A.7b), respectively.

The derivation of the standardized SEP density function in Eq. (A.9) would not be complete without addressing the computation of the first and second moment, $M_1$ and $M_2$, in Eqs. (A.7a) and (A.7b), respectively. As the EP distribution of Subbotin (1923) is symmetric around zero, Eq. (A.8) may be written as follows

$$M_j = 2\int_0^\infty x^j f_{\mathrm{EP}}(x|0,1,\beta)\,\mathrm{d}x, \tag{A.10}$$

and reduces to

$$M_j = 2\omega_\beta \int_0^\infty x^j \exp\bigl(-c_\beta x^{2/(1+\beta)}\bigr)\,\mathrm{d}x, \tag{A.11}$$

for the EP distribution in Eq. (A.1). The above integral appears in integration tables

$$\int_0^\infty x^j \exp(-ax^b)\,\mathrm{d}x = \frac{1}{b}a^{-(j+1)/b}\Gamma\bigl((j+1)/b\bigr), \tag{A.12}$$

where $a = c_\beta$ and $b = 2/(1+\beta)$. If we enter this solution in Eq. (A.11) then we obtain the following closed-form expression for the $j^{\mathrm{th}}$ moment of the EP distribution

$$\begin{aligned} M_j &= 2\omega_\beta \frac{c_\beta^{-(j+1)/(2/(1+\beta))}}{2/(1+\beta)}\Gamma\Bigl((j+1)/\bigl(2/(1+\beta)\bigr)\Bigr) \\ &= \omega_\beta(1+\beta)c_\beta^{-\frac{1}{2}(j+1)(1+\beta)}\Gamma\bigl((j+1)(1+\beta)/2\bigr). \end{aligned} \tag{A.13}$$

For $j = 1$, we yield the first moment, $M_1$, of the EP distribution

$$M_1 = \omega_\beta(1+\beta)c_\beta^{-(1+\beta)}\Gamma(1+\beta). \tag{A.14}$$

[7] With $-1 < \beta \le 1$ the residual norm, $\ell^{2/(1+\beta)}$, ranges between $\ell^1$ and $\ell^\infty$.

We can enter Eqs. (A.2a) and (A.2b) for $\omega_\beta$ and $c_\beta$, respectively, to yield

$$\begin{aligned}
M_1 &= \frac{\Gamma^{1/2}\big(3(1+\beta)/2\big)(1+\beta)}{(1+\beta)\Gamma^{3/2}\big((1+\beta)/2\big)}\left(\left(\frac{\Gamma\big(3(1+\beta)/2\big)}{\Gamma\big((1+\beta)/2\big)}\right)^{1/(1+\beta)}\right)^{-(1+\beta)}\Gamma(1+\beta)\\
&= \frac{\Gamma^{1/2}\big(3(1+\beta)/2\big)}{\Gamma^{3/2}\big((1+\beta)/2\big)}\left(\frac{\Gamma\big(3(1+\beta)/2\big)}{\Gamma\big((1+\beta)/2\big)}\right)^{-1}\Gamma\big(1+\beta\big)\\
&= \frac{\Gamma^{1/2}\big(3(1+\beta)/2\big)}{\Gamma\big(3(1+\beta)/2\big)}\frac{\Gamma\big((1+\beta)/2\big)}{\Gamma^{3/2}\big((1+\beta)/2\big)}\Gamma\big(1+\beta\big), \qquad \text{(A.15)}
\end{aligned}$$

and, thus, the first moment of the EP distribution simplifies to

$$M_1 = \frac{\Gamma(1+\beta)}{\Gamma^{1/2}\big(3(1+\beta)/2\big)\Gamma^{1/2}\big((1+\beta)/2\big)}. \qquad \text{(A.16)}$$

For the second moment, $M_2$, of the EP distribution we must enter, $j = 2$, in Eq. (A.13)

$$M_2 = \omega_\beta(1+\beta)c_\beta^{-\frac{3}{2}(1+\beta)}\Gamma\big(3(1+\beta)/2\big), \qquad \text{(A.17)}$$

After substitution of Eqs. (A.2a) and (A.2b) the above expression simplifies to

$$\begin{aligned}
M_2 &= \frac{\Gamma^{1/2}\big(3(1+\beta)/2\big)(1+\beta)}{(1+\beta)\Gamma^{3/2}\big((1+\beta)/2\big)}\left(\left(\frac{\Gamma\big(3(1+\beta)/2\big)}{\Gamma\big((1+\beta)/2\big)}\right)^{1/(1+\beta)}\right)^{-\frac{3}{2}(1+\beta)}\Gamma\big(3(1+\beta)/2\big)\\
&= \frac{\Gamma^{3/2}\big(3(1+\beta)/2\big)}{\Gamma^{3/2}\big((1+\beta)/2\big)}\left(\frac{\Gamma\big(3(1+\beta)/2\big)}{\Gamma\big((1+\beta)/2\big)}\right)^{-\frac{3}{2}}\\
&= 1. \qquad \text{(A.18)}
\end{aligned}$$

This concludes the derivation.

## Appendix B. The Student *t* likelihood function

The GL and GL$^+$ functions rely on the generalized normal (or EP) distribution to characterize nontraditional residual distributions with different degrees of skew and kurtosis. Scharnagl et al. (2015) replaced the EP distribution of Eq. (7) with a standardized Student's $t$-density

$$f_{\mathrm{ST}}(a|0,1,\nu) = \frac{\Gamma\big((\nu+1)/2\big)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi(\nu-2)}}\left(1+\frac{a^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \qquad \text{(B.1)}$$

where $\mu_a = 0$, $\sigma_a^2 = 1$ and $\nu > 2$ denotes the degrees of freedom. If we combine Eq. (B.1) with the skew density of Eq. (8) we yield the following expression for the standardized skewed Student's $t$ (SST) density (Scharnagl et al., 2015)

$$\begin{aligned}
&f_{\mathrm{SST}}(a|0,1,\nu,\xi)\\
&= \frac{2\sigma_\xi}{(\xi+\xi^{-1})}\frac{\Gamma\big((\nu+1)/2\big)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi a}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi a)}}\right)^2\right)^{-\frac{\nu+1}{2}}, \qquad \text{(B.2)}
\end{aligned}$$

where $\xi > 0$ is the skewness parameter and $\mu_\xi$ and $\sigma_\xi^2$ are shift and scale constants, respectively, which standardize the SST density.

The shift and scale constants, $\mu_\xi$ and $\sigma_\xi^2$, respectively, must satisfy Eqs. (A.7a) and (A.7b)

$$\mu_\xi = M_1(\xi - \xi^{-1}) \qquad \text{(B.3a)}$$

$$\sigma_\xi^2 = (M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2, \qquad \text{(B.3b)}$$

wherein $M_j$ is the $j^{\mathrm{th}}$ absolute moment of the standardized ST density, $f_{\mathrm{ST}}(\cdot|0,1,\nu)$, of Eq. (B.1)

$$M_j = \mathbb{E}(|x^j|) = 2\int_0^\infty x^j f_{\mathrm{ST}}(x|0,1,\nu)\,\mathrm{d}x. \qquad \text{(B.4)}$$

This results in Kirkby et al. (2019)

$$M_j = \frac{\Gamma((j+1)/2)\Gamma\big((\nu-j)/2\big)(\nu-2)^{j/2}}{\sqrt{\pi}\Gamma(\nu/2)}, \qquad \text{(B.5)}$$

and leads to the following expressions for the first two absolute moments of the ST distribution

$$M_1 = \frac{\Gamma\big((\nu-1)/2\big)\sqrt{\nu-2}}{\sqrt{\pi}\Gamma(\nu/2)} \quad \text{and} \quad M_2 = \frac{\Gamma(3/2)\Gamma\big((\nu-2)/2\big)(\nu-2)}{\sqrt{\pi}\Gamma(\nu/2)}. \qquad \text{(B.6)}$$

If we enter the above expressions for $M_1$ and $M_2$ into Eqs. (B.3a) and (B.3b) and simplify and rearrange the resulting Equations we yield (see also Scharnagl et al. 2015)

$$\mu_\xi = \frac{\Gamma\big((\nu-1)/2\big)\sqrt{\nu-2}\,(\xi-\xi^{-1})}{\sqrt{\pi}\,\Gamma(\nu/2)} \qquad \text{(B.7a)}$$

$$\sigma_\xi^2 = -\mu_\xi^2 + \xi^2 + \xi^{-2} - 1. \qquad \text{(B.7b)}$$

The so-obtained shift and scale constants standardize the SST distribution, $\mu_a = 0$ and $\sigma_a^2 = 1$.

The autoregressive model of Eq. (12) may be used to treat serial correlation of the studentized raw residuals, $\underline{\mathbf{e}}(\theta,\delta)$. We can use Eq. (21) to derive the corresponding SST likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta},\nu,\xi,\sigma_\varepsilon^2|\widetilde{\mathbf{y}},\boldsymbol{\sigma}_\epsilon^2) &\simeq \prod_{t=1}^{n}\frac{1}{\sigma_{\epsilon_t}}\frac{1}{\sigma_\varepsilon}f_{\mathrm{SGT}}\big(\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})|\nu,\xi\big)\\
&\simeq \sigma_\varepsilon^{-n}\prod_{t=1}^{n}\frac{2\sigma_\xi\Gamma\big((\nu+1)/2\big)}{\sigma_{\epsilon_t}(\xi+\xi^{-1})\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\\
&\times\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\theta,\delta))}}\right)^2\right)^{-\frac{\nu+1}{2}}, \qquad \text{(B.8)}
\end{aligned}$$

where $\sigma_\varepsilon^2$ equals the variance of the partial residuals in Eq. (13). The above formulation of the SST likelihood function, $L(\boldsymbol{\theta},\nu,\xi,\sigma_\varepsilon^2|\widetilde{\mathbf{y}},\boldsymbol{\sigma}_\epsilon^2)$, assumes prior knowledge of the measurement error variances, $\boldsymbol{\sigma}_\epsilon^2$. We can relax this assumption and infer the $n$ entries of $\boldsymbol{\sigma}_\epsilon^2$ along with the model parameters, $\boldsymbol{\theta}$, using the measurement error function of Eq. (24). The SST likelihood function then becomes

$$\begin{aligned}
L(\boldsymbol{\theta},s_0,\nu,\xi,\sigma_\varepsilon^2|\widetilde{\mathbf{y}}) &\simeq \sigma_\varepsilon^{-n}\prod_{t=1}^{n}\frac{2\sigma_\xi\Gamma\big((\nu+1)/2\big)}{\big(s_0+s_1y_t(\boldsymbol{\theta})\big)(\xi+\xi^{-1})\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\\
&\times\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\theta,\delta))}}\right)^2\right)^{-\frac{\nu+1}{2}}, \qquad \text{(B.9)}
\end{aligned}$$

where $s_1$ equals the zero-point of the residual function, $\hbar(s_1|\boldsymbol{\delta}) = 1 - g(s_1|\boldsymbol{\delta})$, and satisfies the constraint explicated in Eq. (25). The SST log-likelihood, $\mathcal{L}(\boldsymbol{\theta},s_0,\nu,\xi,\sigma_\varepsilon^2|\widetilde{\mathbf{y}})$, now becomes

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta},s_0,\nu,\xi,\sigma_\varepsilon^2|\widetilde{\mathbf{y}}) &\simeq -\frac{n}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2}\sum_{t=1}^{n}\left\{\log\Big(\big(s_0+s_1y_t(\boldsymbol{\theta})\big)^2\Big)\right\}\\
&+ n\log(2) + n\log(\sigma_\xi)\\
&+ n\log\Big(\Gamma\big((\nu+1)/2\big)\Big) - n\log(\xi+\xi^{-1}) - n\log\big(\Gamma(\nu/2)\big) - \frac{n}{2}\log(\pi)\\
&- \frac{n}{2}\log(\nu-2) - \frac{\nu+1}{2}\sum_{t=1}^{n}\left\{\log\left(1+\frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\theta,\delta))}}\right)^2\right)\right\}.
\end{aligned} \qquad \text{(B.10)}$$

This concludes our derivation of the SST log-likelihood function, abbreviated as SL function.

## Appendix C. Special and limiting cases of the SGT distribution

The skewed generalized $t$-distribution of Theodossiou (1998) defines a large family of continuous probability distributions which includes several well-known distributions. These special or limiting cases of the SGT distribution, $f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda,p,q)$, may be derived by fixing one or more of its shape parameters, $\lambda$, $p$ and $q$, to preset values. What follows is a summary of twelve well-known probability distributions.

### C.1. The skewed generalized error distribution

The skewed generalized error (SGE) distribution of Theodossiou (2015) is defined as follows

$$\lim_{q \to \infty} f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda, p, q), \tag{C.1}$$

and yields the following PDF

$$f_{\mathrm{SGE}}(a|\mu, \sigma, \lambda, p) = \frac{p}{2\upsilon\sigma\Gamma(1/p)} \exp\left( -\left( \frac{|a-\mu+m|}{\upsilon\sigma\big(1+\lambda\,\mathrm{sign}(a-\mu+m)\big)} \right)^{p} \right), \tag{C.2}$$

where

$$m = \frac{2^{\frac{2}{p}}\upsilon\sigma\lambda\Gamma(1/2+1/p)}{\sqrt{\pi}} \quad \text{and}$$
$$\upsilon = \sqrt{\frac{\pi\Gamma(1/p)}{\pi(1+3\lambda^2)\Gamma(3/p) - 16^{\frac{1}{p}}\lambda^2\Gamma(1/2+1/p)^2\Gamma(1/p)}}, \tag{C.3}$$

affix the mean, $\mu$, and variance, $\sigma^2$, of the SGE distribution.

### C.2. The generalized Student's *t*-distribution

The generalized Student's $t$-distribution of McDonald and Newey (1988) is defined as follows

$$f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda = 0, p, q), \tag{C.4}$$

and yields the PDF

$$f_{\mathrm{GT}}(a|\mu, \sigma, p, q) = \frac{p}{2\upsilon\sigma q^{1/p} B(1/p, q)} \left( \frac{|a-\mu|^p}{q(\upsilon\sigma)^p} + 1 \right)^{-\left(\frac{1}{p}+q\right)}, \tag{C.5}$$

where

$$\upsilon = \frac{1}{q^{1/p}} \sqrt{\frac{B(1/p, q)}{B(3/p, q-2/p)}}, \tag{C.6}$$

affixes the variance, $\sigma^2$, of the GT distribution.

### C.3. The skewed Student's *t*-distribution

The skewed Student's $t$ (ST)-distribution of Hansen (1994) is defined as follows

$$f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda, p = 2, q), \tag{C.7}$$

and yields the following PDF

$$f_{\mathrm{ST}}(a|\mu, \sigma, \lambda, q)$$
$$= \frac{\Gamma(1/2+q)}{\upsilon\sigma(\pi q)^{1/2}\Gamma(q)} \left( \frac{|x-\mu+m|^2}{q(\upsilon\sigma)^2\big(\lambda\,\mathrm{sign}(x-\mu+m)+1\big)^2} + 1 \right)^{-(\frac{1}{2}+q)}, \tag{C.8}$$

where

$$m = \frac{2\upsilon\sigma\lambda q^{1/2}\Gamma(q-1/2)}{\pi^{1/2}\Gamma(q)} \quad \text{and}$$
$$\upsilon = \frac{1}{q^{1/2}\sqrt{(3\lambda^2+1)(\frac{1}{2q-2}) - \frac{4\lambda^2}{\pi}\left(\frac{\Gamma(q-1/2)}{\Gamma(q)}\right)^2}}, \tag{C.9}$$

affix the mean, $\mu$, and variance, $\sigma^2$, of the ST distribution.

### C.4. The skewed Laplace distribution

The skewed Laplace (SL) distribution is given by

$$\lim_{q \to \infty} f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda, p = 1, q), \tag{C.10}$$

and results in the following PDF

$$f_{\mathrm{SL}}(a|\mu, \sigma, \lambda) = \frac{1}{2\upsilon\sigma} \exp\left( -\left( \frac{|a-\mu+m|}{\upsilon\sigma\big(1+\lambda\,\mathrm{sign}(a-\mu+m)\big)} \right) \right), \tag{C.11}$$

where

$$m = 2\upsilon\sigma\lambda \quad \text{and} \quad \upsilon = \big(2(1+\lambda^2)\big)^{-\frac{1}{2}}, \tag{C.12}$$

affix the mean, $\mu$, and variance, $\sigma^2$, of the SL distribution.

### C.5. The generalized error (normal) distribution

The exponential power (EP) distribution (Subbotin, 1923; Box and Tiao, 1992) is defined as follows

$$\lim_{q \to \infty} f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda = 0, p, q), \tag{C.13}$$

and yields the following PDF

$$f_{\mathrm{EP}}(a|\mu, \sigma, p) = \frac{p}{2\upsilon\sigma\Gamma(1/p)} \exp\left( -\left( \frac{|x-\mu|}{\upsilon\sigma} \right)^{p} \right). \tag{C.14}$$

where

$$\upsilon = \sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}}, \tag{C.15}$$

affixes the variance, $\sigma^2$, of the EP distribution. For $\mu = 0$ and $\sigma = 1$, Eq. (C.14) simplifies to the PDF of the standardized EP distribution in Eq. (7). Indeed, if we substitute Eq. (C.15) into Eq. (C.14) and rearrange the resulting expression we yield

$$f_{\mathrm{EP}}(a|\mu, \sigma, p) = \frac{p\Gamma^{1/2}(3/p)}{2\sigma\Gamma^{3/2}(1/p)} \exp\left( -\left( \frac{\Gamma(3/p)}{\Gamma(1/p)} \right)^{p/2} \left( \frac{|x-\mu|}{\sigma} \right)^{p} \right). \tag{C.16}$$

For $\mu = 0$ and $\sigma = 1$ this expression simplifies to Eq. (7)

$$f_{\mathrm{EP}}(a, p) = \omega_p \exp(-c_p|x|^p), \tag{C.17}$$

with constants $\omega_p$ and $c_p$

$$c_p = \left( \frac{\Gamma(3/p)}{\Gamma(1/p)} \right)^{p/2} \quad \text{and} \quad \omega_p = \frac{p\Gamma^{1/2}(3/p)}{2\Gamma^{3/2}(1/p)}, \tag{C.18}$$

equal to Eqs. (A.2a) and (A.2b), respectively and $p = 2/(1+\beta)$.

### C.6. The skewed normal distribution

The skewed normal (SN) distribution is defined as follows

$$\lim_{q \to \infty} f_{\mathrm{SGT}}(a|\mu, \sigma, \lambda, p = 2, q), \tag{C.19}$$

and has the following PDF

$$f_{\mathrm{SN}}(a|\mu, \sigma, \lambda) = \frac{1}{\upsilon\sigma\sqrt{\pi}} \exp\left( -\left( \frac{|a-\mu+m|}{\upsilon\sigma\big(1+\lambda\,\mathrm{sign}(a-\mu+m)\big)} \right)^{2} \right), \tag{C.20}$$

where

$$m = \frac{2\upsilon\sigma\lambda}{\sqrt{\pi}} \quad \text{and} \quad \upsilon = \sqrt{\frac{2\pi}{(\pi - 8\lambda^2 + 3\pi\lambda^2)}}, \tag{C.21}$$

affix the mean, $\mu$, and variance, $\sigma^2$, of the SN distribution.

### C.7. The Student's *t*-distribution

The Student's $t$-distribution (Student, 1908) is given by

$$f_{\mathrm{SGT}}(a|\mu = 0, \sigma = 1, \lambda = 0, p = 2, q), \tag{C.22}$$

and yields the PDF

$$f_{\mathrm{T}}(a|\nu) = \frac{\Gamma\big((\nu+1)/2\big)}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \left( 1 + \frac{x^2}{\nu} \right)^{-\frac{\nu+1}{2}}. \tag{C.23}$$

### C.8. The skewed Cauchy distribution

The skewed Cauchy distribution is defined as follows

$$f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda,p=2,q=1), \tag{C.24}$$

and yields the following PDF

$$f_{\mathrm{SC}}(a|\mu,\sigma,\lambda) = \frac{1}{\sigma\pi}\left(\frac{|x-\mu|^2}{\sigma^2\big(\lambda\,\mathrm{sign}(x-\mu)+1\big)^2}+1\right)^{-1}. \tag{C.25}$$

### C.9. The Laplace distribution

The Laplace (1774) distribution is given by

$$\lim_{q\to\infty} f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda=0,p=1,q), \tag{C.26}$$

which results in the following PDF

$$f_{\mathrm{L}}(a|\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|a-\mu|}{\sigma}\right), \tag{C.27}$$

where $\sigma$ is the scale parameter of the Laplace distribution.

### C.10. The uniform distribution

The uniform distribution is given by

$$\lim_{\substack{p\to\infty\\ q\to\infty}} f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda=0,p,q), \tag{C.28}$$

and results in the following PDF

$$f_{\mathrm{U}}(a) = \begin{cases} \dfrac{1}{2\upsilon\sigma} & \text{if } |a-\mu| < \upsilon\sigma \\ 0 & \text{otherwise.} \end{cases} \tag{C.29}$$

The standard uniform distribution is obtained by setting $\mu = (a+b)/2$, $\upsilon = 1$, and $\sigma = (b-a)/2$.

### C.11. The normal distribution

The Gaussian distribution is defined as follows

$$\lim_{q\to\infty} f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda=0,p=2,q), \tag{C.30}$$

and yields the familiar PDF

$$f_{\mathrm{N}}(a|\mu,\sigma) = \frac{1}{\upsilon\sigma\sqrt{\pi}}\exp\left(-\left(\frac{a-\mu}{\upsilon\sigma}\right)^2\right), \tag{C.31}$$

where $\upsilon = \sqrt{2}$ affixes the variance, $\sigma^2$, of the normal distribution.

### C.12. The Cauchy-Lorentz distribution

The Cauchy-Lorentz (CL) distribution of Poisson (1824) is given by

$$f_{\mathrm{SGT}}(a|\mu,\sigma,\lambda=0,p=2,q=1), \tag{C.32}$$

and results in the following PDF

$$f_{\mathrm{CL}}(a|\mu,\sigma) = \frac{1}{\sigma\pi}\left(\left(\frac{x-\mu}{\sigma}\right)^2+1\right)^{-1}. \tag{C.33}$$

This concludes Appendix C.

## Appendix D. The SGT likelihood with normally distributed residuals

If the partial residuals, $\varepsilon_t(\boldsymbol{\theta},\boldsymbol{\delta})$, are believed to be normally distributed, then we must set, $\lambda = 0$, $p = 2$ and $q \to \infty$. Unfortunately, we cannot submit an infinite value of $q$ to Eqs. (39) and/or (41) as this will induce numerical problems. A pragmatic remedy is to enter a large value of $q$ instead, say, the default value of $q = 10^{10}$ in Table 4. This will do in practice. We must still demonstrate the theoretical equivalence of the SGT log-likelihood function with $\lambda = 0$, $p = 2$ and $q \to \infty$ and the NL function of Eq. (27).

To support limit analysis, we print below the SGT log-likelihood function of Eq. (41) and label the three terms that require limit analysis

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},s_0,\lambda,p,q,\boldsymbol{\Phi}_2|\widetilde{\mathbf{y}}) \simeq{}& -\frac{n}{2}\log(\sigma_\varepsilon^2) - \sum_{t=1}^{n}\Big\{\log\big(|s_0+s_1y_t(\boldsymbol{\theta})|\big)\Big\} + n\log(p)\\ & - n\log(2) - \underbrace{n\log(\kappa_{\lambda pq})}_{\text{(i)}} - \underbrace{n\log\big(B(1/p,q/p)\big)}_{\text{(ii)}}\\ & \underbrace{-\frac{q+1}{p}\sum_{t=1}^{n}\left\{\log\left(1+\left|\frac{\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})+\mu_{\lambda pq}}{\kappa_{\lambda pq}\big(1+\lambda\,\mathrm{sign}(\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})+\mu_{\lambda pq})\big)}\right|^p\right)\right\}}_{\text{(iii)}}.\end{aligned} \tag{D.1}$$

We use the labels, (i), (ii) and (iii) in the analysis below.

For $\lambda = 0$, $p = 2$ and $q \to \infty$, the shift constant, $\mu_{02\infty}$, in Eq. (35a) amounts to zero. For the scale constant, $\kappa_{02\infty}$, in Eq. (35b) we yield Eq. (D.2) (see Box II). We can take advantage of the following limit of the beta function

$$\lim_{b\to\infty} B(a,b) = \Gamma(a)\,b^{-a}, \tag{D.3}$$

to simplify the expression of the scale constant, $\kappa_{02\infty}$, to yield

$$\begin{aligned}\kappa_{02q} &= \sqrt{\frac{\Gamma(1/2)\,(q/2)^{-1/2}}{\Gamma(3/2)\,\big((q-2)/2\big)^{-3/2}}} && \text{if } q\to\infty\\ &= \sqrt{\frac{2\Gamma(1/2)\,(q/2)^{-1/2}}{\Gamma(1/2)\,\big((q-2)/2\big)^{-3/2}}} && \text{if } q\to\infty\\ &= \sqrt{\frac{2(q/2)^{-1/2}}{(q/2)^{-3/2}}} && \text{if } q\to\infty\\ &= \sqrt{q} && \text{if } q\to\infty.\end{aligned} \tag{D.4}$$

Thus, for $q \to \infty$, term (i) in Eq. (D.1) is equal to $n\log(\sqrt{q})$.

The limit of the beta function in Eq. (D.3) can also be applied to term (ii) of Eq. (D.1). We yield

$$\begin{aligned}\lim_{q\to\infty} n\log\big(B(1/2,q/2)\big) &= n\log\big(\Gamma(1/2)(q/2)^{-1/2}\big)\\ &= \frac{n}{2}\log(\pi) - \frac{n}{2}\log(q/2)\\ &= \frac{n}{2}\log(2\pi) - n\log(\sqrt{q})\end{aligned} \tag{D.5}$$

Thus, in the limit of $q \to \infty$, term (ii) of Eq. (D.1) may be replaced with $\frac{n}{2}\log(2\pi) - n\log(\sqrt{q})$.

This leaves us with term (iii) of the SGT log-likelihood function in Eq. (D.1). The following limit is well-known in elementary calculus

$$\lim_{x\to\infty}\left(1+\frac{a}{x}\right)^x = \exp(a). \tag{D.6}$$

If we take the natural logarithm of both sides of this limit and then divide by $p$, we yield the corollary

$$\lim_{x\to\infty}\frac{x}{p}\log\left(1+\frac{a}{x}\right) = \frac{a}{p}. \tag{D.7}$$

For $q = x$, $p = 2$, and $a = \underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2$, the left-hand side reduces to term (iii) of Eq. (D.1), but without the sum operator. As a result, we yield

$$\lim_{q\to\infty}\frac{q+1}{p}\sum_{t=1}^{n}\left\{\log\left(1+\frac{\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2}{q}\right)\right\}$$

$$
\begin{aligned}
\kappa_{02\infty} &= \lim_{q\to\infty} \frac{B(1/2,q/2)}{\sqrt{(1+3\times 0^2)B(1/2,q/2)B\big(3/2,(q-2)/2\big) - 4\times 0^2 \times B\big(1,(q-1)/2\big)^2}} \\
&= \lim_{q\to\infty} \sqrt{\frac{B(1/2,q/2)}{B\big(3/2,(q-2)/2\big)}} \\
&= \sqrt{\frac{\lim_{q\to\infty} B(1/2,q/2)}{\lim_{q\to\infty} B\big(3/2,(q-2)/2\big)}}. \qquad \text{(D.2)}
\end{aligned}
$$

**Box II.**

$$
\begin{aligned}
&= \frac{\underline{\varepsilon}_1(\boldsymbol{\theta},\boldsymbol{\delta})^2}{2} + \frac{\underline{\varepsilon}_2(\boldsymbol{\theta},\boldsymbol{\delta})^2}{2} + \cdots + \frac{\underline{\varepsilon}_n(\boldsymbol{\theta},\boldsymbol{\delta})^2}{2} \\
&= \frac{1}{2}\sum_{t=1}^{n} \underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2, \qquad \text{(D.8)}
\end{aligned}
$$

and, thus, in the limit of $q \to \infty$, term (iii) of Eq. (D.1) simplifies to half the sum of squares of the standardized partial residuals.

If we replace terms (i), (ii) and (iii) in Eq. (D.1) with their counterparts of Eqs. (D.4), (D.5) and (D.8) we yield the SGT log-likelihood function, $\mathcal{L}(\boldsymbol{\theta}, s_0, \boldsymbol{\Phi}_2|\widetilde{\mathbf{y}}, \lambda = 0, p = 2, q \to \infty)$, for normally-distributed partial residuals

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}, s_0, \boldsymbol{\Phi}_2|\widetilde{\mathbf{y}}, \lambda = 0, p = 2, q \to \infty) \\
&\simeq -\frac{n}{2}\log(\sigma_\varepsilon^2) - \sum_{t=1}^{n}\big\{\log\big(|s_0 + s_1 y_t(\boldsymbol{\theta})|\big)\big\} + n\log(2) - n\log(2) \\
&\quad - n\log(\sqrt{q}) - \Big(\frac{n}{2}\log(2\pi) - n\log(\sqrt{q})\Big) - \frac{1}{2}\sum_{t=1}^{n}\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2. \qquad \text{(D.9)}
\end{aligned}
$$

This expression of the SGT log-likelihood function reduces to the NL function of Eq. (27)

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}, s_0, \boldsymbol{\Phi}_2|\widetilde{\mathbf{y}}, \lambda = 0, p = 2, q \to \infty) \\
&\simeq -\frac{n}{2}\log(2\pi) - \sum_{t=1}^{n}\big\{\log\big(|s_0 + s_1 y_t(\boldsymbol{\theta})|\big)\big\} \\
&- \frac{n}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2}\sum_{t=1}^{n}\underline{\varepsilon}_t(\boldsymbol{\theta},\boldsymbol{\delta})^2, \qquad \text{(D.10)}
\end{aligned}
$$

with variance of the partial residuals, $\sigma_\varepsilon^2$, which is derived from Eq. (13). Thus, the SGT distribution is the key ingredient of a large family of likelihood functions, which includes the normal likelihood function.

This concludes the derivation.

## Appendix E. Algorithmic recipe for the predictive distributions

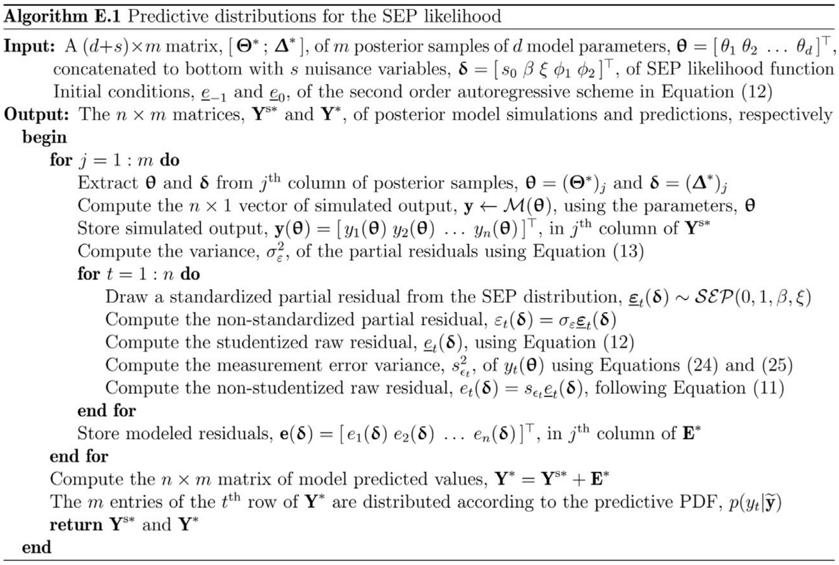
Algorithm E.1 presents a step-by-step recipe for determining the predictive distribution, $p(y_t|\widetilde{\mathbf{y}})$, of Eq. (3) from the posterior realizations sampled with DREAM Suite. This recipe is specifically written for the SEP likelihood of the GL$^+$ function, but easy to adapt to the SL and UL functions of Eqs. (31) and (41), respectively.

The algorithmic recipe uses as input arguments, (i) the $(d+s)\times m$ matrix, $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$, of $m$ posterior samples of $d$ model parameters, $\boldsymbol{\theta} = [\,\theta_1\ \theta_2\ \dots\ \theta_d\,]^\top$, concatenated vertically with the $s$ nuisance variables, $\boldsymbol{\delta} = [s_0\ \ \beta\ \ \xi\ \ \phi_1\ \ \phi_2]^\top$, of the SEP likelihood function and (ii) the initial conditions, $\underline{e}_{-1}$ and $\underline{e}_0$, of the second order autoregressive scheme in Eq. (12), and returns $n\times m$ matrices of posterior model simulations, $\mathbf{Y}^{\mathrm{s}*}$, and predictions, $\mathbf{Y}^*$.

Predictive percentiles, $y^*_{t,\alpha/2}$ and $y^*_{t,1-\alpha/2}$, are readily derived from the $m$ entries, $y^*_{tj} \in \mathbb{R}^{n\times m}$, in the $t^{\text{th}}$ row of $\mathbf{Y}^*$

$$
\int_{y^*_{t,\alpha/2}}^{y^*_{t,1-\alpha/2}} p(y_t|\widetilde{\mathbf{y}}) = p(y^*_{t,\alpha/2} \le y^*_{tj} \le y^*_{t,1-\alpha/2}|\widetilde{\mathbf{y}}) = 1-\alpha \qquad \forall\ j = (1,2,\dots,m), \qquad \text{(E.1)}
$$

by truncating $\alpha/2$ of the left and right tails of the predictive distribution, $p(y_t|\widetilde{\mathbf{y}})$. This must be repeated for all $t = (1,2,\dots,n)$ and produces two $n\times 1$-vectors, $\mathbf{y}^*_{\alpha/2}$ and $\mathbf{y}^*_{1-\alpha/2}$, respectively, which together define



```
Algorithm E.1 Predictive distributions for the SEP likelihood
Input: A $(d{+}s){\times}m$ matrix, $[\,\boldsymbol{\Theta}^*\,;\,\boldsymbol{\Delta}^*\,]$, of $m$ posterior samples of $d$ model parameters, $\boldsymbol{\theta} = [\,\theta_1\ \theta_2\ \dots\ \theta_d\,]^\top$,
    concatenated to bottom with $s$ nuisance variables, $\boldsymbol{\delta} = [\,s_0\ \beta\ \xi\ \phi_1\ \phi_2\,]^\top$, of SEP likelihood function
    Initial conditions, $\underline{e}_{-1}$ and $\underline{e}_0$, of the second order autoregressive scheme in Equation (12)
Output: The $n \times m$ matrices, $\mathbf{Y}^{\mathrm{s}*}$ and $\mathbf{Y}^*$, of posterior model simulations and predictions, respectively
  begin
    for $j = 1 : m$ do
      Extract $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ from $j^{\text{th}}$ column of posterior samples, $\boldsymbol{\theta} = (\boldsymbol{\Theta}^*)_j$ and $\boldsymbol{\delta} = (\boldsymbol{\Delta}^*)_j$
      Compute the $n \times 1$ vector of simulated output, $\mathbf{y} \leftarrow \mathcal{M}(\boldsymbol{\theta})$, using the parameters, $\boldsymbol{\theta}$
      Store simulated output, $\mathbf{y}(\boldsymbol{\theta}) = [\,y_1(\boldsymbol{\theta})\ y_2(\boldsymbol{\theta})\ \dots\ y_n(\boldsymbol{\theta})\,]^\top$, in $j^{\text{th}}$ column of $\mathbf{Y}^{\mathrm{s}*}$
      Compute the variance, $\sigma_\varepsilon^2$, of the partial residuals using Equation (13)
      for $t = 1 : n$ do
        Draw a standardized partial residual from the SEP distribution, $\underline{\varepsilon}_t(\boldsymbol{\delta}) \sim \mathcal{SEP}(0,1,\beta,\xi)$
        Compute the non-standardized partial residual, $\varepsilon_t(\boldsymbol{\delta}) = \sigma_\varepsilon \underline{\varepsilon}_t(\boldsymbol{\delta})$
        Compute the studentized raw residual, $\underline{e}_t(\boldsymbol{\delta})$, using Equation (12)
        Compute the measurement error variance, $s^2_{\epsilon_t}$, of $y_t(\boldsymbol{\theta})$ using Equations (24) and (25)
        Compute the non-studentized raw residual, $e_t(\boldsymbol{\delta}) = s_{\epsilon_t}\underline{e}_t(\boldsymbol{\delta})$, following Equation (11)
      end for
      Store modeled residuals, $\mathbf{e}(\boldsymbol{\delta}) = [\,e_1(\boldsymbol{\delta})\ e_2(\boldsymbol{\delta})\ \dots\ e_n(\boldsymbol{\delta})\,]^\top$, in $j^{\text{th}}$ column of $\mathbf{E}^*$
    end for
    Compute the $n \times m$ matrix of model predicted values, $\mathbf{Y}^* = \mathbf{Y}^{\mathrm{s}*} + \mathbf{E}^*$
    The $m$ entries of the $t^{\text{th}}$ row of $\mathbf{Y}^*$ are distributed according to the predictive PDF, $p(y_t|\widetilde{\mathbf{y}})$
    return $\mathbf{Y}^{\mathrm{s}*}$ and $\mathbf{Y}^*$
  end
```

**Table F.1**
Matrix of Pearson's linear correlation coefficients of the scoring rules and performance metrics of the UL, SL, GL$^+$ and NL functions of Table 8.

| | Unit | Scoring rules, $\mathcal{S}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | | Performance metrics, $\mathcal{P}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LS | CRPS | SS | IS | RLBL | CV | *C* | *W* |
| LS | log(d/mm) | **1.00** | 0.32 | 0.36 | −0.03 | 0.03 | −0.19 | 0.68 | 0.53 |
| CRPS | mm/d | | **1.00** | 0.47 | 0.67 | −0.33 | −0.11 | 0.05 | 0.71 |
| SS | – | | | **1.00** | −0.11 | −0.82 | − 0.73 | −0.04 | 0.10 |
| IS | mm/d | | | | **1.00** | 0.02 | 0.41 | −0.17 | 0.65 |
| RLBL | – | | | | | **1.00** | 0.82 | 0.49 | 0.25 |
| CV | – | | | | | | **1.00** | 0.19 | 0.36 |
| *C* | – | | | | | | | **1.00** | 0.54 |
| *W* | mm/d | | | | | | | | **1.00** |

**Table F.2**
Matrix of Spearman's rank correlation coefficients of the scoring rules and performance metrics of the UL, SL, GL$^+$ and NL functions of Table 8.

| | Unit | Scoring rules, $\mathcal{S}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | | Performance metrics, $\mathcal{P}(\mathbb{F},\widetilde{\mathbf{y}})$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LS | CRPS | SS | IS | RLBL | CV | *C* | *W* |
| LS | log(d/mm) | **1.00** | 0.42 | 0.27 | 0.03 | 0.02 | −0.04 | 0.73 | 0.49 |
| CRPS | mm/d | | **1.00** | 0.24 | 0.62 | −0.18 | 0.07 | 0.28 | 0.87 |
| SS | – | | | **1.00** | 0.02 | −0.76 | −0.68 | −0.16 | 0.10 |
| IS | mm/d | | | | **1.00** | 0.08 | 0.27 | −0.13 | 0.69 |
| RLBL | – | | | | | **1.00** | 0.91 | 0.48 | 0.14 |
| CV | – | | | | | | **1.00** | 0.35 | 0.33 |
| *C* | – | | | | | | | **1.00** | 0.48 |
| *W* | mm/d | | | | | | | | **1.00** |

the $100(1-\alpha)\%$ prediction uncertainty. If we admit instead to Eq. (E.1) the $n \times m$ matrix $\mathbf{Y}^{S*}$ of posterior model simulations then we yield prediction percentiles due to parameter uncertainty only.

## Appendix F. Correlation of scoring rules and performance metrics

To determine whether the scoring rules and/or performance metrics measure independent information about the discharge forecast distribution, we compute pairwise correlation coefficients of the different scoring rules and performance metrics of Table 8.

## References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21 (10), 5293–5313. http://dx.doi.org/10.5194/hess-21-5293-2017.

Alexander, C., Coulon, M., Han, Y., Meng, X., 2022. Evaluating the discrimination ability of proper multi-variate scoring rules. Ann. Oper. Res. http://dx.doi.org/10.1007/s10479-022-04611-9.

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. http://dx.doi.org/10.5194/hess-23-2147-2019.

Anderson, R.L., 1941. Distribution of the serial correlation coefficients. Ann. Math. Stat. 8 (1), 1–13.

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. http://dx.doi.org/10.1029/2000WR900363.

Bayes, T., 1763. An essay toward solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.. Philos. Trans. R. Soc. Lond. 53, 370–418. http://dx.doi.org/10.1098/rstl.1763.0053.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1), 18–36. http://dx.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. http://dx.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. http://dx.doi.org/10.1016/S0022-1694(01)00421-8.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Methodol. 26 (2), 211–243. http://dx.doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Box, G.E.P., Tiao, G.C., 1992. Bayesian Inference in Statistical Analysis. John Wiley & Sons, New York, NY, USA, http://dx.doi.org/10.1002/9781118033197.

Boyle, D.P., 2001. Multicriteria Calibration of Hydrological Models. (Ph.D. thesis). Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ.

ter Braak, C.J.F.T., 2006. A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. 16 (3), 239–249. http://dx.doi.org/10.1007/s11222-006-8769-1.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 1–3.

Brooks, S., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7, 434–455.

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. http://dx.doi.org/10.5194/hess-24-3951-2020.

Cosslett, S.R., 1983. Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51 (3), 765–782.

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. J. Amer. Statist. Assoc. 91, 883–904.

Damsleth, E., El-Shaarawi, A.H., 1989. ARMA models with double exponentially distributed noise. J. R. Stat. Soc. Ser. B Stat. Methodol. 51 (1), 61–69.

Dawid, A.P., 2007. The geometry of proper scoring rules. Ann. Inst. Statist. Math. 59 (1), 77–93. http://dx.doi.org/10.1007/s10463-006-0099-8.

Evin, G., Kavetski, D., Thyer, M., Kuczera, G., 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resour. Res. 49 (7), 4518–4524. http://dx.doi.org/10.1002/wrcr.20284.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. 50 (3), 2350–2375. http://dx.doi.org/10.1002/2013WR014185.

Fenicia, F., Kavetski, D., Reichert, P., Albert, C., 2018. Signature-domain calibration of hydrological models using approximate Bayesian computation: Empirical analysis of fundamental properties. Water Resour. Res. 54 (6), 3958–3987. http://dx.doi.org/10.1002/2017WR021616.

Fernandez, C., Steel, M.J.F., 1998. On Bayesian modeling of fat tails and skewness. J. Amer. Statist. Assoc. 93, 359–371.

Fisher, R.A., 1934. Statistical Methods for Research Workers, fifth ed. In: Biological monographs and manuals, no. 5, Oliver and Boyd, Edinburgh, UK.

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. http://dx.doi.org/10.1029/95WR03723.

Friedman, D., 1983. Effective scoring rules for probabilistic forecasts. Manage. Sci. 29 (4), 447–454. http://dx.doi.org/10.1287/mnsc.29.4.447.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data Analysis, thirdrd ed Chapman & Hall/CRC, Boca Raton.

Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 7, 457–511. http://dx.doi.org/10.1214/ss/1177011136.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. vol. 91, Oxford University Press, pp. 169–193.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold-and quantile-weighted scoring rules. J. Bus. Econom. Statist. 29 (3), 411–422.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc. Ser. B Stat. Methodol. 14 (1), 107–114.

Good, I.J., 1971. Discussion of "measuring information and uncertainty" by R. J. Buehler. In: Godambe, V.P., Sprott, D.A. (Eds.), Foundations of Statistical Inference. Holt, Rinehardt and Winston, Toronto, CA, pp. 337–339.

Greenberg, E., 2013. Introduction To Bayesian Econometrics, second ed. Cambridge University Press, New York.

Hamill, T.M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev. 129 (3), 550–560. http://dx.doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hansen, B.E., 1994. Autoregressive conditional density estimation. Internat. Econom. Rev. 35, 705–730.

Hansen, J.V., McDonald, J.B., Turley, R.S., 2006. Partially adaptive robust estimation of regression models and applications. European J. Oper. Res. 170 (1), 132–143. http://dx.doi.org/10.1016/j.ejor.2004.06.008.

Hernández-López, M.R., Francés, F., 2017. Bayesian joint inference of hydrological and generalized error models with the enforcement of total laws. Hydrol. Earth Syst. Sci. Discuss. 2017, 1–40. http://dx.doi.org/10.5194/hess-2017-9.

Hollaway, M.J., Beven, K.J., Benskin, C.M.W.H., Collins, A.L., Evans, R., Falloon, P.D., Forber, K.J., Hiscock, K.M., Kahana, R., Macleod, C.J.A., Ockenden, M.C., Villamizar, M.L., Wearing, C., Withers, P.J.A., Zhou, J.G., Barber, N.J., Haygarth, P.M., 2018. A method for uncertainty constraint of catchment discharge and phosphorus load estimates. Hydrol. Process. 32 (17), 2779–2787. http://dx.doi.org/10.1002/hyp.13217.

Hürlimann, W., 2012. On non-Gaussian AR(1) inflation modeling. J. Stat. Econ. Methods 1 (1), 93–109.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions, second ed. In: Wiley Series in Probability and Statistics, vol. 2, John Wiley & Sons, New York, NY, USA.

Kirkby, J.L., Nguyen, D., Nguyen, D., 2019. Moments of student's t-distribution: A unified approach. SSRN http://dx.doi.org/10.2139/ssrn.3497188.

Krüger, F., Lerch, S., Thorarinsdottir, T., Gneiting, T., 2021. Predictive inference based on Markov chain Monte Carlo output. Internat. Statist. Rev. 89 (2), 274–301. http://dx.doi.org/10.1111/insr.12405.

Kuczera, G., 1982. On the relationship between the reliability of parameter estimates and hydrologic time series data used in calibration. Water Resour. Res. 18 (1), 146–154. http://dx.doi.org/10.1029/WR018i001p00146.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86. http://dx.doi.org/10.1214/aoms/1177729694.

Laloy, E., Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing. Water Resour. Res. 48 (1), http://dx.doi.org/10.1029/2011WR010608.

Laplace, P.S., 1774. Mémoire sur la probabilité des causes par les évènements. Mémoires de L'académie Royale Des Sci. Presentés Par Divers Savan 6, 621–656.

Lee, P.M., 2012. Bayesian Statistics: An Introduction, forth ed. Wiley, Chichester, England.

Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F., Gneiting, T., 2017. Forecaster's dilemma: extreme events and forecast evaluation. Statist. Sci. 32 (1), 106–127. http://dx.doi.org/10.1214/16-STS588.

Li, M., Wang, Q.J., Bennett, J.C., Robertson, D.E., 2016. Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. Hydrol. Earth Syst. Sci. 20, 3561–3579. http://dx.doi.org/10.5194/hess-20-3561-2016.

Lu, D., Ye, M., Hill, M.C., 2012. Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. Water Resour. Res. 48 (9), 1087–1096. http://dx.doi.org/10.1029/2011WR011289.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Manage. Sci. 22, 1087–1096.

McDonald, J.B., Newey, W.K., 1988. Partially adaptive estimation of regression models via the generalized t distribution. Econom. Theory 4, 428–457.

McInerney, D., Kavetski, D., Thyer, M., Lerat, J., Kuczera, G., 2019. Benefits of explicit treatment of zero flows in probabilistic hydrological modeling of ephemeral catchments. Water Resour. Res. 55 (12), 11035–11060. http://dx.doi.org/10.1029/2018WR024148.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53 (3), 2199–2239. http://dx.doi.org/10.1002/2016WR019168.

McMillan, H.K., Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty. Hydrol. Process. 29 (7), 1873–1882. http://dx.doi.org/10.1002/hyp.10419.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 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. http://dx.doi.org/10.5194/hess-19-209-2015.

Oliveira, D.Y., Vrugt, J.A., 2022. The treatment of uncertainty in diagnostic model evaluation: A probabilistic description of streamflow records. Water Resour. Res. http://dx.doi.org/10.1029/2022WR032263.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J. Hydrol. 303 (1), 290–306. http://dx.doi.org/10.1016/j.jhydrol.2004.08.026.

Poisson, S.D., 1824. Sur la probabilite des resultats moyens des observations. Connaissance Des Tems Pour L'An 1827 273–302.

Raftery, A.E., Lewis, S., 1992. How many iterations in the gibbs sampler? In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. vol. 91, Oxford University Press, pp. 763–773.

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. Water Resour. Res. 45 (10), http://dx.doi.org/10.1029/2009WR007814.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resour. Res. 47 (11), http://dx.doi.org/10.1029/2011WR010643.

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. http://dx.doi.org/10.5194/hess-17-4831-2013.

Sadegh, M., Vrugt, J.A., Xu, C., Volpi, E., 2015. The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(ABC)}$. Water Resour. Res. 51 (11), 9207–9231. http://dx.doi.org/10.1002/2014WR016805.

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), http://dx.doi.org/10.1029/2009WR008933.

Shannon, C.E., 1948a. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shannon, C.E., 1948b. A mathematical theory of communication. Bell Syst. Tech. J. 27 (4), 623–656. http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Shuford, E.H., Albert, A., Massengill, H.E., 1966. Admissible probability measurement procedures. Psychometrika 31, 125–145.

Smith, P., Beven, K.J., Tawn, J.A., 2008. Informal likelihood measures in model assessment: Theoretic development and investigation. Adv. Water Resour. 31 (8), 1087–1100. http://dx.doi.org/10.1016/j.advwatres.2008.04.012.

Smith, T.J., Marshall, L.A., 2008. Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques. Water Resour. Res. 44 (12), http://dx.doi.org/10.1029/2007WR006705.

Smith, T., Marshall, L., Sharma, A., 2015. Modeling residual hydrologic errors with Bayesian inference. J. Hydrol. 528, 29–37. http://dx.doi.org/10.1016/j.jhydrol.2015.05.051.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., Sisson, S., 2010. Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. Water Resour. Res. 46 (12), http://dx.doi.org/10.1029/2010WR009514.

Sorooshian, S., Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases. Water Resour. Res. 16 (2), 430–442. http://dx.doi.org/10.1029/WR016i002p00430.

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. http://dx.doi.org/10.1029/wr021i009p01421.

Steinschneider, S., McCrary, R., Wi, S., Mulligan, K., Mearns, L.O., Brown, C., 2015. Expanded decision-scaling framework to select robust long-term water-system plans under hydroclimatic uncertainties. J. Water Resour. Plan. Manag. 141 (11), 04015023. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000536.

Student, 1908. The probable error of a mean. Biometrika 6, 1–25.

Subbotin, M.T., 1923. On the law of frequency of error. Mat. Sb. 31, 296–301.

Tasker, G.D., 1980. Hydrologic regression with weighted least squares. Water Resour. Res. 16 (6), 1107–1113. http://dx.doi.org/10.1029/wr016i006p01107.

Theodossiou, P., 1998. Financial data and the skewed generalized t distribution. Manage. Sci. 44 (12), 1650–1661. http://dx.doi.org/10.1287/mnsc.44.12.1650.

Theodossiou, P., 2015. Skewed generalized error distribution of financial assets and options pricing. Multinat. Finance J. 19, 223–266. http://dx.doi.org/10.17578/19-4-1.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W., Srikanthan, S., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resour. Res. 45 (12), http://dx.doi.org/10.1029/2008WR006825.

Van Geert, P., 1994. Dynamic Systems of Development: Change Between Complexity and Chaos. Harvester, New York, NY.

Volpi, E., Schoups, G., Firmani, G., Vrugt, J.A., 2017. Sworn testimony of the model evidence: Gaussian mixture importance (GAME) sampling. Water Resour. Res. 53 (7), 6133–6158. http://dx.doi.org/10.1002/2016WR020167.

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316. http://dx.doi.org/10.1016/j.envsoft.2015.08.013.

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. http://dx.doi.org/10.1016/j.jhydrol.2018.02.026.

Vrugt, J.A., ter Braak, C.J.F., 2011. DREAM$_{(D)}$: an adaptive Markov chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. Hydrol. Earth Syst. Sci. 15, 3701–3713. http://dx.doi.org/10.5194/hess-15-3701-2011.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44 (12), http://dx.doi.org/10.1029/2007WR006720.

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. http://dx.doi.org/10.1515/IJNSNS.2009.10.3.273.

Vrugt, J.A., Sadegh, M., 2013a. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. Water Resour. Res. 49 (7), 4335–4345. http://dx.doi.org/10.1002/wrcr.20354.

Winkler, R.L., Muññoz, J., Cervera, J.L., Bernardo, J.M., Blattenberger, G., Kadane, J.B., Lindley, D.V., Murphy, A.H., Oliver, R.M., Ríı́os-Insua, D., 1996. Scoring rules and the evaluation of probabilities. Test 5 (1), 1–60. http://dx.doi.org/10.1007/BF02562681.

Yang, J., Reichert, P., Abbaspour, K.C., 2007. Bayesian uncertainty analysis in distributed hydrologic modeling: A case-study in the Thur River basin (Switzerland). Water Resour. Res. 43 (W10401), http://dx.doi.org/10.1029/2006WR005497.

Zeckhauser, R., Thompson, M., 1970. Linear regression with non-normal error terms. Rev. Econ. Stat. 52 (3), 280–286.