Hydrol. Earth Syst. Sci., 23, 2147–2172, 2019
https://doi.org/10.5194/hess-23-2147-2019

Hydrology and Earth System Sciences
Open Access
EGU

A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation

Lorenz Ammann[1,2], **Fabrizio Fenicia**[1], **and Peter Reichert**[1,2]

[1]Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dubendorf, Switzerland
[2]Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

Correspondence: Lorenz Ammann (lorenz.ammann@eawag.ch)

Received: 27 July 2018 – Discussion started: 13 August 2018
Revised: 25 March 2019 – Accepted: 26 March 2019 – Published: 30 April 2019

Abstract. The widespread application of deterministic hydrological models in research and practice calls for suitable methods to describe their uncertainty. The errors of those models are often heteroscedastic, non-Gaussian and correlated due to the memory effect of errors in state variables. Still, residual error models are usually highly simplified, often neglecting some of the mentioned characteristics. This is partly because general approaches to account for all of those characteristics are lacking, and partly because the benefits of more complex error models in terms of achieving better predictions are unclear. For example, the joint inference of autocorrelation of errors and hydrological model parameters has been shown to lead to poor predictions. This study presents a framework for likelihood functions for deterministic hydrological models that considers correlated errors and allows for an arbitrary probability distribution of observed streamflow. The choice of this distribution reflects prior knowledge about non-normality of the errors. The framework was used to evaluate increasingly complex error models with data of varying temporal resolution (daily to hourly) in two catchments. We found that (1) the joint inference of hydrological and error model parameters leads to poor predictions when conventional error models with stationary correlation are used, which confirms previous studies; (2) the quality of these predictions worsens with higher temporal resolution of the data; (3) accounting for a non-stationary autocorrelation of the errors, i.e. allowing it to vary between wet and dry periods, largely alleviates the observed problems; and (4) accounting for autocorrelation leads to more realistic model output, as shown by signatures such as the flashiness index. Overall, this study contributes to a better description of residual errors of deterministic hydrological models.

1 Introduction

Deterministic hydrological models are widely applied in research and decision-making processes. The quantification of their associated uncertainties is therefore an important task with high relevance for the scientific learning process, as well as for operational decisions with respect to water management. The total output uncertainty of those models is a combination of (i) propagated input uncertainty (e.g. Sun et al., 2000; Kavetski et al., 2003; Bárdossy and Das, 2008); (ii) model structural errors (e.g. Butts et al., 2004), which can be attributed to aggregation and parameterisation; and (iii) parameter uncertainty (e.g. Freer et al., 1996; Wagener et al., 2001). When performing inference, (iv) observation errors are an additional source of uncertainty, which arise for example due to errors in rating curves (e.g. Kuczera and Franks, 2002). The sources (i–iv) usually result in residual errors of predicted streamflow observations with the following characteristics:

- *Non-normality.* Model residuals are seldom well represented by a normal distribution with constant mean and variance. Instead, residuals are typically heteroscedastic (increasing with streamflow), right-skewed due to non-negativity of streamflow and characterized by excess kurtosis (fat tails) (e.g. Schoups and Vrugt, 2010).

- *Autocorrelation.* Several sources of error cause memory effects. Such sources are inadequacy of model structure, errors in internal states of the model (Kavetski et al., 2003) or missed rainfall events, which can have an effect on the residuals several days after the event has occurred (e.g. Beven and Westerberg, 2011).

- *Non-stationarity.* Model residuals can have very different characteristics in time. For example, during wet periods dominated by rainfall, errors are generally less correlated than during dry periods (Yang et al., 2007). Schaefli et al. (2007) find that residuals are less correlated during high flows than during low flows in a glacierised alpine catchment.

- *Unequally spaced observations.* Observations do not always take place at fixed time intervals. Particularly for water quality, volume-proportional sampling strategies are generally preferable to fixed-time strategies (e.g. Schleppi et al., 2006). These strategies generate observations at unequal time intervals. Another cause of unequal observation intervals is missing data.

Various studies have investigated error models that consider correlation, heteroscedasticity and non-normality of errors of deterministic hydrological models. A typical approach, which is also applied in this study, is to describe total output uncertainty in a lumped way (e.g. Schoups and Vrugt, 2010; McInerney et al., 2017). Another group of approaches distinguishes among the different sources of total uncertainty such as input, parametric and output measurement uncertainty (e.g. Kavetski et al., 2006; Renard et al., 2010). The latter approach is conceptually desirable, but it can lead to identifiability problems and it is computationally very intensive due to the required propagation of errors through the model. For many applications we need a computationally cheaper approach that can be achieved with a lumped model. It is the goal of this paper to contribute to the improvement of these lumped approaches. Current approaches to describe total output uncertainty in a lumped way differ in if, and how, they deal with the various characteristics of residual errors mentioned above. Some of the most common approaches are the following:

- Heteroscedasticity is often considered in weighted least-squares error models by parameterising the variance of the normal distribution as a function of the streamflow (Thyer et al., 2009; Evin et al., 2013; Bertuzzo et al., 2013). Another common approach is to apply transformations such as Box–Cox to the observed and modelled streamflow time series and formulate a model for the residuals of the transformed time series (e.g. Bates and Campbell, 2001; Del Giudice et al., 2013; McInerney et al., 2017). However, this transformation affects several properties of the residuals simultaneously, including heteroscedasticity, skewness and kurtosis.

- Typically, residual errors are represented as a stationary process. The issue of stationarity has been the subject of recent debate (Milly et al., 2008; Montanari and Koutsoyiannis, 2014). Focusing on streamflow dynamics, an example of representing non-stationarity of residual errors is shown in the study of Yang et al. (2007), who distinguish between wet and dry periods by applying a continuous autoregressive process with different parameters for the wet and the dry periods to the Box–Cox transformed residuals.

- A probabilistic model to deal with unequally spaced data was proposed by Duan et al. (1988). A more natural formulation is to adopt a continuous-time formulation of the autoregressive model, such as an Ornstein–Uhlenbeck process (OU process; e.g. Kloeden and Platen, 1995; Yang et al., 2007).

- Non-negativity of streamflow can be addressed by truncating the error probability density function so that it does not extend to negative streamflow. This leads to zero probability for zero streamflow, which may not always be adequate. The truncation approach is seldom followed, and in most applications the truncation occurs "in prediction only" (McInerney et al., 2017).

Residual error models are usually highly simplified, in the sense that they do not account for all the above-mentioned characteristics of these errors. In particular, residual error models seldom go beyond using "variance stabilisation" techniques such as Box–Cox transformations. The widespread use of relatively simple error models is due to several reasons. In our opinion, the following are the most important.

First, there is a lack of general approaches that can deal with all the above-mentioned characteristics of error models simultaneously. One general error model that can accommodate various characteristics is the probabilistic model proposed by Schoups and Vrugt (2010), which can deal with residual errors that are correlated, heteroscedastic and non-Gaussian with varying degrees of kurtosis and skewness. They do this by describing the errors with an autoregressive process with a skew exponential power (SEP) rather than a normal distribution for the innovations. However, their approach is shown to produce unrealistically large predictive uncertainties caused by the application of the autoregressive process to non-standardised residuals (Evin et al., 2013). Scharnagl et al. (2015) attempt to address this issue by applying an autoregressive process to the standardised residuals of a soil moisture model, using a skewed Student's t distribution to describe the probability density of the innovations of the autoregressive process. However, with this approach they experience problematic inference behaviour and biased results similar to those mentioned by Evin et al. (2013). Furthermore, while the conventional approach of using normal innovations for the errors leads to a normal marginal of (potentially transformed) streamflow, non-normal innovations lead to marginal streamflow distributions which are generally not available in closed form. An explicit marginal distribution of streamflow (Krzysztofowicz, 2002) facilitates scientific communication and discussion, since hydrologists are generally more familiar with streamflow than with Box–Cox trans-

formation parameters or distributions of the innovations of residuals.

Second, there is limited guidance to the choice of a particular error model for a given application. In the past, the choice has been generally ad hoc, with limited justification. Only recently, there has been more systematic comparison and testing which has resulted in some general recommendations. For example, McInerney et al. (2017) compare various residual error schemes, including standard and weighted least squares, the Box–Cox transformation (with fixed and calibrated power parameter) and the log-sinh transformation using data from 23 catchments and concluded that Box–Cox has on average the best behaviour.

Third, previous experience has shown that more realistic error models, which are more complex, do not always result in better predictions. The additional parameters of some of the more complex error models were found to have undesirable interactions with the parameters of the hydrological model, leading to unrealistic parameter values and poor predictions. For example, particularly in dry catchments, accounting for autocorrelation produces worse predictions than omitting it (Schoups and Vrugt, 2010; Evin et al., 2013). To circumvent such problems, Evin et al. (2014) recommend that autoregressive parameters are inferred sequentially, that is, after having estimated all other parameters of the hydrological and of the error model. Similarly, many uncertainty analysis techniques are applied for fixed hydrological parameters, avoiding the re-calibration of hydrological models (e.g. Montanari and Brath, 2004). The joint inference of hydrological and error model parameters remains conceptually preferable, as it recognises potential interactions between parameters. The conditions under which this can be achieved remain poorly understood.

Fourth, the potential advantages of more complex error models are under-appreciated by the hydrological community. For relatively simple uncertainty analysis, like the plotting of uncertainty bands around hydrographs, the use of simplified error models may appear justified. However, there are several applications that go beyond this task, and for which a simplified error model may lead to poor results. For example, assuming uncorrelated errors may lead to unrealistic extrapolations (Del Giudice et al., 2013) or too-strong short-term fluctuations, which have a large effect on hydrograph signatures that are sensitive to noise, such as the flashiness index (Baker et al., 2004; Fenicia et al., 2018). The ability to correctly represent signatures is not only important for conceptual reasons, but also for practical purposes such as in signature-based model calibration.

The goals of this study are the following:

1. Develop a flexible framework for likelihood functions for hydrological models that accounts for the following major characteristics of their errors: non-normality (heteroscedasticity, skewness and excess kurtosis), autocorrelation, non-stationarity regarding wet and dry periods, unequally spaced observation time points, and non-negativity of streamflow.

2. Use the flexible framework to do controlled experiments by varying some of the assumptions and by performing joint inference of a hydrological model with error models of increasing complexity. Investigate the effect of the various assumptions on the quality of the predictive distributions. In particular, with case studies in two catchments, we investigate the following questions:

   (a) Can we confirm previous findings about the problems related to joint inference of hydrological and error model parameters?

   (b) What are the causes of the problems encountered in joint inference of hydrological and error model parameters?

   (c) Can we improve the joint inference by introducing non-stationarity by allowing the autoregressive parameter to change between wet and dry periods?

   (d) Does the consideration of autocorrelation lead to more realistic predictions (e.g. in terms of better representation of hydrograph signatures such as the flashiness index)?

   (e) Can parameters controlling the shape of the distribution of the errors be inferred jointly with the hydrological model parameters to account for non-normality?

The paper is structured as follows. The theoretical framework for the probabilistic model, corresponding to Goal 1, is presented in Sect. 2.1 and the performance metrics used to evaluate it are described in Sect. 2.4. Section 3 describes the case study set-up used to carry out the necessary investigations for Goal 2. The case study is based on two catchments (Sect. 3.1), one hydrological bucket model (Sect. 3.2) and three different time step sizes (daily, 6-hourly and hourly). The results of those investigations are presented in Sect. 4 and discussed in Sect. 5. Section 6 lists the main conclusions and sketches potential directions for future research.

2 Methods

2.1 Probabilistic framework

Suppose we choose the distribution D_Q to describe the probability of observing streamflow Q, given the model output Q_{det} (see Fig. 1). We believe that this is a natural place to start the derivation of a probabilistic framework for hydrological models, since it enables us to communicate and discuss the basic assumptions in a space that is most familiar to hydrological modellers: the space of streamflow. Note the major difference to transformation-based approaches (e.g.

Bates and Campbell, 2001; Del Giudice et al., 2013; McInerney et al., 2017) and approaches that use non-normal innovations of the stochastic process (Schoups and Vrugt, 2010; Scharnagl et al., 2015), both of which lead to D_Q not being readily available in closed form. In particular, discussing the possible distribution of streamflow given the output of a hydrological model is easier than discussing Box–Cox transformation parameters or the distribution of the innovations of the model errors. Providing explicit control over D_Q therefore facilitates the formulation of the model based on prior knowledge resulting from past experience of hydrologists in units they are familiar with. Wani et al. (2019) present another approach in which D_Q at subsequent output time steps is accessed through copulas.

We assume that D_Q is parameterised by Q_{det} and some error model parameters $\boldsymbol{\psi}$, i.e. $Q(t) \sim D_Q(Q_{\text{det}}(t, \boldsymbol{\theta}), \boldsymbol{\psi})$, where $\boldsymbol{\theta}$ are the parameters of the deterministic hydrological model. This implies that the observed streamflow at different time points can be described by different distributions (e.g. with varying mean and standard deviation), but these distributions belong to the same parametric family. The distribution D_Q may extend to negative values. In this case, the integrated probability of negative values is assigned to the probability of observing a streamflow of zero. This leads to

$$p_{D_Q(Q_{\text{det}}, \boldsymbol{\psi})}(Q) = \begin{cases} f_{D_Q(Q_{\text{det}}, \boldsymbol{\psi})}(Q) & \text{if } Q > 0, \\ F_{D_Q(Q_{\text{det}}, \boldsymbol{\psi})}(0) & \text{if } Q = 0, \\ 0 & \text{if } Q < 0, \end{cases} \tag{1}$$

where f_{D_Q} and F_{D_Q} are the density and cumulative distribution function of D_Q, respectively, and p is a probability density for $Q > 0$ and a discrete probability for $Q = 0$. Note that Eq. (1) reflects our prior knowledge that $Q \geq 0$ when dealing with non-tidal rivers. If the distribution chosen for D_Q is limited to positive support, either by choosing a distribution with positive support or by truncating at zero, only the first case in Eq. (1) applies and we get zero probability for $Q = 0$. This is a common approach that is fully covered by the presented framework. However, especially in ephemeral catchments, a finite probability for $Q = 0$ might be desirable (Smith et al., 2010). This can be achieved by choosing a distribution D_Q that extends to negative values. Equation (1) then assigns the negative tail to $Q = 0$. If correlation is absent or neglected, Eq. (1) can be applied at each time step and the likelihood function is simply the product of those mutually independent terms.

Accounting for temporal correlation requires some additional conceptualisations. Consider the transformation function

$$\eta_{\text{trans}}(Q, Q_{\text{det}}, \boldsymbol{\psi}) = F^{-1}_{\text{N}(0,1)}\left(F_{D_Q(Q_{\text{det}}, \boldsymbol{\psi})}(Q)\right), \tag{2}$$

which transforms the streamflow, Q, via its assumed marginal distribution, D_Q, which is dependent on the model output, Q_{det}. If the distributional assumptions for D_Q are correct, the result of this transformation is a standard normally distributed variable. Applying Eq. (2) to a time series of streamflow, $Q(t_i)$, leads to a time series of transformed streamflows:

$$\eta(t_i) = \eta_{\text{trans}}(Q(t_i), Q_{\text{det}}(t_i), \boldsymbol{\psi}), \tag{3}$$

where t_i are the time points of interest for inference or prediction. Note that, if the distributional assumptions about D_Q hold at all points in time, $\eta(t_i)$ are a sample from a standard normal distribution, except for the lower tail, which can be lighter due to the truncation at zero at each individual time step.

To describe autocorrelation in the deviations of Q from Q_{det}, we assume that the corresponding time series of η are discrete-time results of a continuous-time autoregressive process:

$$\eta(t_i)|\eta(t_{i-1}) \sim \text{N}\left(\eta(t_{i-1}) \exp\left(-\frac{t_i - t_{i-1}}{\tau(t_i)}\right), \sqrt{1 - \exp\left(-2\frac{t_i - t_{i-1}}{\tau(t_i)}\right)}\right) \tag{4}$$

where N is the normal distribution and the first and the second argument is the mean and the standard deviation, respectively. This so-called Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930) has a standard normal asymptotic distribution and a characteristic correlation time, $\tau(t_i)$, that is assumed to be constant over the interval $[t_{i-1}, t_i]$.

In summary, to transfer information between time points, we transform the distribution D_Q at time t_{i-1} to a standard normal distribution η_{i-1} according to Eq. (2), advance η_{i-1} to η_i according to Eq. (4), and transform η_i back to D_Q at time t_i.

Note that, for a constant time step $\Delta t = t_i - t_{i-1}$, Eq. (4) becomes

$$\eta(t_i|t_{i-1}) \sim \text{N}\left(\eta(t_{i-1})\phi, \sqrt{1 - \phi^2}\right), \tag{5}$$

with

$$\phi = \exp(-\frac{\Delta t}{\tau}) \quad \text{or} \quad \tau = -\frac{\Delta t}{\ln(\phi)}. \tag{6}$$

This is a discrete-time AR(1) process with autoregression coefficient ϕ and white noise variance $1 - \phi^2$. The formulation of a continuous-time autoregressive process with evaluation at discrete time points allows us to apply it to non-equidistant time series. One advantage of this formulation is that it combines autocorrelation with the possibility to easily deal with missing data, which is considerably more difficult when using the fixed-time version in Eq. (5). Note that the continuous-time formulation assumes that η can be described well by an autoregressive process of first order, where in fact higher orders have been observed (Kuczera, 1983; Bates and Campbell, 2001). Nonetheless, the first-order approximation has been used often throughout hydrological literature.

In order to formulate the probability of the streamflow Q, we used Eqs. (1) to (4) to derive the following conditional probabilities for $Q(t_i)$ given $Q(t_{i-1})$ (see Appendix A for the full derivation).

$$
\begin{aligned}
&\text{If } Q(t_{i-1}) > 0: \\
&p_i\big(Q(t_i)|Q(t_{i-1}), \boldsymbol{\theta}, \boldsymbol{\psi}\big) \\
&= \begin{cases} f_{D_Q(Q_{\mathrm{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi})}(Q(t_i)) \dfrac{f_{\mathrm{N}\left(\eta(t_{i-1})\exp\left(-\frac{t_i-t_{i-1}}{\tau(t_i)}\right),\sqrt{1-\exp\left(-2\frac{t_i-t_{i-1}}{\tau(t_i)}\right)}\right)}(\eta(t_i))}{f_{\mathrm{N}(0,1)}(\eta(t_i))} \\ \quad \text{if } Q(t_i) > 0, \\ F_{\mathrm{N}\left(\eta(t_{i-1})\exp\left(-\frac{t_i-t_{i-1}}{\tau(t_i)}\right),\sqrt{1-\exp\left(-2\frac{t_i-t_{i-1}}{\tau(t_i)}\right)}\right)}(\eta(t_i)) \\ \quad \text{if } Q(t_i) = 0. \end{cases} \\
&\text{If } Q(t_{i-1}) = 0: \\
&p_i\big(Q(t_i)|Q(t_{i-1}), \boldsymbol{\theta}, \boldsymbol{\psi}\big) \\
&= \begin{cases} f_{D_Q\left(Q_{\mathrm{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\right)}\big(Q(t_i)\big) & \text{if } Q(t_i) > 0, \\ F_{D_Q\left(Q_{\mathrm{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\right)}(0) & \text{if } Q(t_i) = 0. \end{cases}
\end{aligned} \tag{7}
$$

Note that p is a probability density (denoted by f) if $Q(t_i) > 0$, and an integrated, discrete probability (denoted by F) if $Q(t_i) = 0$. Note also that η in Eq. (7) is calculated with Eq. (3) and depends on Q and $Q_{\mathrm{det}}(\boldsymbol{\theta})$. Furthermore, Eq. (7) reduces to Eq. (1) for $(t_i - t_{i-1})/\tau \to \infty$, i.e. if the characteristic correlation time is short compared to the length of the time step.

The likelihood is then obtained by building the product of the conditional probabilities in Eq. (7) and by substituting the observations, Q_{obs}, for Q:

$$
\begin{aligned}
&f_{\mathrm{L}}\big(Q_{\mathrm{obs}}(t_0),\, Q_{\mathrm{obs}}(t_1),\, \ldots,\, Q_{\mathrm{obs}}(t_n)|\boldsymbol{\theta}, \boldsymbol{\psi}\big) = \\
&\quad p_{D_Q(Q_{\mathrm{det}}(t_0,\boldsymbol{\theta}),\boldsymbol{\psi})}\big(Q_{\mathrm{obs}}(t_0)\big) \\
&\quad \prod_{i=1}^{n} p_i\big(Q_{\mathrm{obs}}(t_i)|Q_{\mathrm{obs}}(t_{i-1}), \boldsymbol{\theta}, \boldsymbol{\psi}\big).
\end{aligned} \tag{8}
$$

Note that the first term on the right hand side of Eq. (8) can be calculated with Eq. (1), since it is not conditional on the previous time step.

Zeger and Brookmeyer (1986) and Hannachi (2012) formulated a likelihood that allows the memory of an autoregressive processes to be kept during time periods with censored data. This concept can be transferred to the case of zero streamflow. It has a conceptual advantage over Eq. (7), especially when dealing with intermittent data with frequent periods with observations of zero that can be shorter than the characteristic correlation length, like for example in the case of precipitation (Hannachi, 2012). Depending on a catchment's low-pass filtering effect, streamflow is expected to have fewer but longer continuous periods of zero and non-zero data compared to precipitation. Consequently, the memory of the process given by Eq. (4) is likely to vanish during a zero streamflow period of typical length, reducing the benefit of keeping the correlation during those periods. Therefore, the cost of numerically solving integrals, the dimension of which is proportional to the length of the zero streamflow period (Hannachi, 2012), outweighs the conceptual benefits with respect to this application. The approach by Zeger and Brookmeyer (1986) might be highly relevant in other hydrological applications, however.

2.2 Error models

As a basis for subsequent applications, we set D_Q to the skewed Student's t distribution (Fig. 1), which is obtained by transforming the conventional Student's t distribution according to Fernandez and Steel (1998). This approach of skewing has been used in a previous study on error models (Schoups and Vrugt, 2010), albeit in a different setting. Thus, we introduce two error model parameters: γ, defining the degree of skewness, and d_{f}, the degrees of freedom as a measure for the kurtosis. The skewed Student's t distribution reduces to the normal distribution for $\gamma = 1$ and $d_{\mathrm{f}} \to \infty$. Two assumptions are tested to centre D_Q at Q_{det}:

$$E[D_Q] = Q_{\mathrm{det}}(t), \tag{9a}$$

$$\mathrm{mode}(D_Q) = Q_{\mathrm{det}}(t), \tag{9b}$$

i.e. we either assign the expected value or the highest probability density of D_Q to Q_{det}. A third alternative would be to set the median of D_Q equal to Q_{det}. By testing the two options in Eq. (9), we include the lowest and the highest value; the third option would be a compromise between the two and was not included in the study. If not indicated otherwise, the assumption in Eq. (9a) was used. The results obtained with Eq. (9b) can be found in Appendix B.

The standard deviation of D_Q is parameterised as follows:

$$\sigma_{D_Q}(t) = a Q_0 \left(\frac{Q_{\mathrm{det}}(t)}{Q_0}\right)^c + b Q_0. \tag{10}$$

Note that skewing a distribution with the approach developed by Fernandez and Steel (1998) changes its standard deviation; $\sigma_{D_Q}(t)$ is the standard deviation of D_Q after skewing. Other parameterisations of σ_{D_Q} are in principle possible; see McInerney et al. (2017) for a theoretical correspondence with transformation approaches. McInerney et al. (2017) have shown that transformation approaches with a first-order correspondence to $c = 0.8$ or $c = 0.5$ can lead to more reliable and precise predictions than those corresponding to $c = 1$. To limit the scope of the analysis, and to maintain comparability to previous studies (Thyer et al., 2009; Schoups and Vrugt, 2010; Evin et al., 2013), we set c equal to 1. Note that the parameters a and b become dimensionless (and therefore more universal) by including a reference streamflow, Q_0, that corresponds to the mean of the observations: $Q_0 = \overline{Q}_{\mathrm{obs}}$. Thus, a accounts for the variable and b for the constant contributions to the total standard deviation.

Table 1. Overview of the error models applied in this study, their assumptions regarding correlation and the distribution of streamflow and their corresponding parameters (SKT: skewed Student's t distribution, $\times$: fitted).

Error model	Distribution	Correlation	a	b	γ	d_f	τ_{min}	τ_{max}
E1	Gaussian	none	$\times$	$\times$	1	∞	0	0
E2	Gaussian	constant	$\times$	$\times$	1	∞	$= \tau_{max}$	$\times$
E3*	Gaussian	non-stationary, partially fitted	$\times$	$\times$	1	∞	0	$\times$
E3a*	Gaussian	non-stationary, fitted	$\times$	$\times$	1	∞	$\times$	$\times$
E4*	SKT	non-stationary, partially fitted	$\times$	$\times$	$\times$	$\times$	0	$\times$
E4a*	SKT	non-stationary, fitted	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$

If * is appended to the name of the error model, a smoothed version of $P_{err}(t)$ (moving average of window size 5 h) was used in Eq. (11).

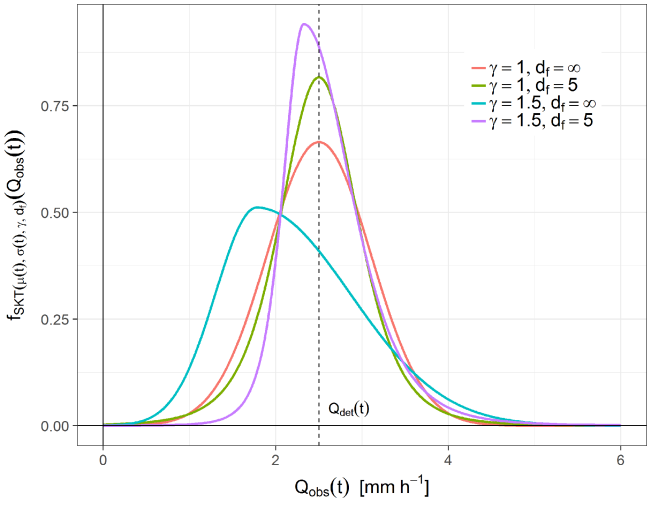

Figure 1. Example of skewed Student's t distributions with $E[D_Q] = Q_{det}(t) = 2.5\,\mathrm{mm\,h^{-1}}$ and standard deviation $\sigma_{D_Q}(t) = 0.6\,\mathrm{mm\,h^{-1}}$ for different values of skewness, γ, and degrees of freedom, d_f.

Table 1 lists the error models applied in this study, together with their underlying assumptions. E1 is included as a reference case; it is based on the assumption of uncorrelated heteroscedastic errors with a normal distribution. These assumptions, with the exception of heteroscedasticity and the treatment of $Q_{obs} = 0$, are identical to those made when maximising the Nash–Sutcliffe efficiency for example, or, equivalently, minimising the squared residuals. Error model E2 represents a conventional approach to considering autocorrelation. In the case of equally spaced time steps, it is similar to the error model applied by Evin et al. (2013) for example, who assume that the rescaled errors follow an AR(1) process with a standard normal marginal distribution. One difference between the two approaches is, again, the treatment of cases where $Q_{obs} = 0$. In error model E3, we additionally account for the fact that τ might be time-dependent. The following formula for τ is used in those cases:

$$\tau(t) = \begin{cases} \tau_{min} & \text{if } P_{err}(t) > 0, \\ \tau_{max} & \text{otherwise}, \end{cases} \tag{11}$$

where P_{err} is the precipitation used as an input for the error model. In E3, τ_{min} is fixed at 0, while in E3a, it is fitted. P_{err} was either equal to the recorded precipitation, P, or, in the case of hourly resolution in the Maimai catchment, smoothed with a moving average of window size 5 h. This was done to prevent frequent jumps between τ_{min} and τ_{max} during precipitation events, and to be more robust with respect to potential time lags between observed precipitation and streamflow. Note that, if such time lags were excessively large, they would have to be considered in Eq. (11). Since in the Murg catchment smoothing did not change the results substantially, $P_{err} = P$ applies there. Thus, error model E3a (or E3) can be seen as a mixture of E1 and E2, in the sense that τ alternates between periods of high and low (or no) correlation. Finally, E4 relaxes the assumption of normality for D_Q; we use a skewed Student's t distribution, inferring the degrees of freedom and the skewness. Again, E4a denotes the version where τ_{min} is inferred.

2.3 Inference and prediction

Consider that for any practical case of inference or prediction, we will have a finite series of time points of interest $(t_0, t_1, \ldots, t_n)$ and a corresponding time series of streamflow $\boldsymbol{Q} = (Q(t_0), Q(t_1), \ldots, Q(t_n))$ or, in analogy, $\boldsymbol{Q}_{det}$ and $\boldsymbol{Q}_{obs}$. When performing inference, the parameters of the hydrological model, $\boldsymbol{\theta}$, are estimated jointly with the parameters of the error model, $\boldsymbol{\psi}$, by evaluating the likelihood function (Eq. 8) according to the following procedure:

1. Given a suggested parameter vector $\boldsymbol{\theta}$, evaluate the deterministic hydrological model, $\boldsymbol{Q}_{det}$, for all time points.

2. Using $\boldsymbol{\psi}$ and $\boldsymbol{Q}_{det}$, calculate the likelihood in Eq. (8).

As the likelihood (Eq. 8) is available in closed form for a given output of the hydrological model, like in many common likelihood functions in hydrology, we do Bayesian inference based on standard MCMC sampling of the posterior. The affine-invariant ensemble sampler by Foreman-Mackey et al. (2013) is used for this purpose. It uses the so-called "stretch move" to propose a new value for a point in parameter space based on other members of the ensemble. The ensemble size consists of 100 walkers in this study and convergence is assessed visually. A full posterior sample consists of 10 000 model evaluations after successful convergence.

For prediction, stochastic realisations of model output are obtained by inverting Eq. (2):

$$Q_{\text{trans}}(\eta,\, Q_{\text{det}},\, \boldsymbol{\psi}) = F^{-1}_{D_Q(Q_{\text{det}},\, \boldsymbol{\psi})}\big(F_{\text{N}(0,\,1)}(\eta)\big), \tag{12}$$

and applying the following procedure to produce a single stochastic streamflow realisation $\boldsymbol{Q}_j$:

1. Randomly draw a parameter vector $(\boldsymbol{\theta}, \boldsymbol{\psi})_j$ from the posterior sample.
2. Using $\boldsymbol{\theta}_j$, evaluate the deterministic hydrological model to obtain $\boldsymbol{Q}_{\text{det},\,j}$ for all time points.
3. Using $\boldsymbol{\tau}_j \in \boldsymbol{\psi}_j$ and Eq. (4), produce a stochastic realisation of an OU process, $\boldsymbol{\eta}_j$, with a standard normal marginal distribution.
4. Use $\boldsymbol{\psi}_j$ and $\boldsymbol{Q}_{\text{det},\,j}$, determined in steps 1 and 2, to transform $\boldsymbol{\eta}_j$ into a stochastic realisation of streamflow, $\boldsymbol{Q}_j$, with Eq. (12).

Note that a simulation with the hydrological model requires some additional input like precipitation and potential evapotranspiration data (Sect. 3.1), which is assumed to be known also for the prediction period. In a synthetic case study, we could successfully verify the consistency of the implemented likelihood and sampling functions (see the Supplement).

2.4 Evaluation criteria

How can the performance of empirical error models, such as those presented in this study, be quantified? We argue that the performance of an error model in joint inference with a hydrological model should be judged according to the following criteria: (a) good reproduction of observed dynamic fluctuations by individual model realisations, (b) good overall predictive marginal distribution of streamflow, and (c) small absolute deviance between model output and observations. The flashiness index (Sect. 2.4.1) is an indicator for (a). The reliability and the relative spread of the predictive distribution (Sect. 2.4.2 and 2.4.3, respectively) are used as an indicator for (b). The Nash–Sutcliffe efficiency (Sect. 2.4.4) and the relative error in cumulative streamflow (Sect. 2.4.5) cover (c). In addition to those performance metrics, we calculated the Kullback–Leibler divergence (Kullback and Leibler, 1951) of the marginal posterior parameter distributions from the prior according to the method proposed by Boltz et al. (2007).

2.4.1 Flashiness index

The function to calculate the flashiness index (Baker et al., 2004) is given by the following:

$$I(\boldsymbol{Q}) = \frac{\sum_{i=1}^{n} |Q(t_i) - Q(t_{i-1})|}{\sum_{i=1}^{n} Q(t_i)}, \tag{13}$$

where $\boldsymbol{Q} = (Q(t_0),\, Q(t_1),\, \ldots,\, Q(t_n))$. Let $\widehat{x}$ denote the quantity x that is related to the hydrological parameter values at the maximum posterior density. The flashiness index is calculated for the observations, $I_{\text{F, obs}} = I(\boldsymbol{Q}_{\text{obs}})$; the output of the deterministic hydrological model, $\widehat{I}_{\text{F, det}} = I(\widehat{\boldsymbol{Q}}_{\text{det}})$; and the individual stochastic realisations of the predictive streamflow sample, $I_{\text{F}} = \text{median}(I(\boldsymbol{Q}_j))$. I_{F} is sensitive to the amount of autocorrelation in a streamflow time series, as well as the height of the peaks of $\boldsymbol{Q}_{\text{det}}$ (since $\boldsymbol{Q}_j$ depends on $\boldsymbol{Q}_{\text{det}}$).

2.4.2 Reliability

Reliability is defined similarly to McInerney et al. (2017), as follows:

$$\Xi_{\text{reli}} = 1 - \frac{2}{n+1} \sum_{i=0}^{n} |F_{Q(t_i)}(Q_{\text{obs}}(t_i)) - F_{\zeta}(F_{Q(t_i)}(Q_{\text{obs}}(t_i)))|, \tag{14}$$

where $\boldsymbol{\zeta} = \{F_{Q(t_i)}(Q_{\text{obs}}(t_i)) | i \in \mathbb{N},\, 0 \le i \le n\}$, F_{ζ} is the empirical cumulative distribution function of $\boldsymbol{\zeta}$ and $F_{Q(t_i)}$ is the empirical cumulative distribution function of the predicted streamflow at time t_i. Ξ_{reli} can take values in the interval [0, 1], where larger values of Ξ_{reli} correspond to better reliability and unity means perfect reliability. It measures the degree to which the observations are consistent with being a sample of the predictive distribution. Since comparison happens in the uniform space, the influence of heavy outliers on Ξ_{reli} is limited. Note that we use the complement of the reliability measure proposed by McInerney et al. (2017), in order to allow for a more intuitive interpretation (larger values mean larger reliability).

2.4.3 Relative spread

The relative spread is an indicator for the width of the predictive distributions over all time points, and was proposed by McInerney et al. (2017) as follows:

$$\Omega_{\text{spread}} = \frac{\sum_{i=0}^{n} \sigma_Q(t_i)}{\sum_{i=0}^{n} Q_{\text{obs}}(t_i)}, \tag{15}$$

where $\sigma_Q(t_i)$ is the standard deviation of the predictive distribution at time point t_i calculated from the ensemble of all

stochastic predictions at that point in time. $\Omega_{\text{spread}} \in \mathbb{R}^+$, and small values of Ω_{spread} indicate precise predictions or small predictive uncertainty. The smaller the predictive uncertainty, the better the quality of the underlying model, given that the predictions are not overconfident. While McInerney et al. (2017) use the name "precision" for Ω_{spread}, we believe that "relative spread" is a more appropriate term considering its actual meaning.

2.4.4 Nash–Sutcliffe efficiency

The Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970), $E_{\text{N, f}}$ (f for function), is defined as follows:

$$E_{\text{N, f}}(\boldsymbol{Q},\ \boldsymbol{Q}_{\text{obs}}) = 1 - \frac{\sum_{i=0}^{n}(Q(t_i) - Q_{\text{obs}}(t_i))^2}{\sum_{i=0}^{n}(Q_{\text{obs}}(t_i) - \overline{Q}_{\text{obs}})^2}, \tag{16}$$

where $\boldsymbol{Q} = (Q(t_0),\ Q(t_1),\ \ldots,\ Q(t_n))$. It is used in this study to assess the output of the hydrological model at the maximum posterior parameter density, $\widehat{E}_{\text{N, det}} = E_{\text{N, f}}(\widehat{\boldsymbol{Q}}_{\text{det}},\ \boldsymbol{Q}_{\text{obs}})$, as well as the stochastic simulations, $E_{\text{N}} = \text{median}(E_{\text{N, f}}(\boldsymbol{Q}_j,\ \boldsymbol{Q}_{\text{obs}}))$. It is used as a rough measure of how well two hydrographs correspond to each other, primarily with the goal of identifying very poorly fitting hydrographs. It is known to be sensitive to errors in high flows (Legates and McCabe, 1999), which can be of particular practical interest. Therefore it complements the other measures, which are less informative with respect to errors in high flows.

2.4.5 Relative error in total cumulative streamflow

As a measure of systematic over- or under-prediction of streamflow, we calculate the relative error in total cumulative streamflow:

$$\Delta(\boldsymbol{Q},\ \boldsymbol{Q}_{\text{obs}}) = \frac{\sum_{i=0}^{n} Q_{\text{obs}}(t_i) - Q(t_i)}{\sum_{i=0}^{n} Q_{\text{obs}}(t_i)}. \tag{17}$$

It is calculated with respect to the model output based on the parameter values at the maximum posterior density; $\widehat{\Delta}_{Q,\text{det}} = \Delta(\widehat{\boldsymbol{Q}}_{\text{det}},\ \boldsymbol{Q}_{\text{obs}})$, as well as for the ensemble of individual stochastic simulations: $\Delta_Q = \text{median}(\Delta(\boldsymbol{Q}_j,\ \boldsymbol{Q}_{\text{obs}}))$. Note that, contrary to McInerney et al. (2017), Δ_Q is the median error of all the individual hydrograph realisations, not the error of the average hydrograph.

3 Case study set-up

3.1 Catchments and data

The probabilistic framework developed in Sect. 2.1 was tested in two case study sites, the Murg and the Maimai catchments, which are described in this section. The Murg river flows through a hilly headwater catchment in a temperate climate with a size of 80 km^2 in northeastern Switzerland. Some key hydrological summary statistics are listed in Table 2. Land use is predominantly agricultural (50 %), with forested headwaters (30 %) and a considerable proportion of urban areas (10 %). The mean elevation is 652 m a.s.l., spanning from 466 to 1035 m a.s.l. Streamflow peaks can be quite sharp, especially for small events, in which baseflow conditions are reached again within just a few hours. This is potentially due to impervious areas being drained directly into the river. The data consist of hourly averages of streamflow, precipitation and potential evapotranspiration from January 1995 to December 2002. Calibration was performed in the first 5 years (January 1995–December 1999) and validation in the consecutive 3 years (January 2000–December 2002). Streamflow data are courtesy of the Swiss Federal Office for the Environment (FOEN). Precipitation and potential evapotranspiration are based on meteorological data (MeteoSwiss, 2018) and were processed by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), with the preprocessing tools of PREVAH (Viviroli et al., 2009).

The Maimai experimental catchments are a set of small headwater catchments with a long history of hydrological research. They are located on a deeply incised hillslope on the South Island of New Zealand. The area is forested and the climate is considerably more humid than in the Murg catchment (Table 2). The site was chosen for this study due to its homogeneous characteristics and relatively simple hydrological response, which make it very suited for model evaluation and testing (e.g. Seibert and McDonnell, 2002). We use hourly data recorded in 1985–1987 in the M8 experimental catchment, the most intensely studied of the Maimai catchments. It has an area of ca. 7 ha with steep (34°) slopes. The reader is referred to Brammer and McDonnell (1996) for a more detailed description of the characteristics of the M8 and the other experimental catchments. This study does not attempt to make a significant contribution to the understanding of the hillslope processes in the Maimai catchment (see McGlynn et al., 2002, for an extensive overview). Calibration was performed based on data from January 1985–December 1986, and validation during January–December 1987. The data were kindly provided by Jeffrey McDonnell.

While the resolution of the original data was hourly, we produced data sets with 6-hourly and daily resolution by aggregation for both catchments. This set-up allows us to systematically investigate the effect of the temporal resolution of the data on the joint inference of hydrological and error model parameters. This could contribute to the identification of the cause of previously encountered problems in joint inference (Goal 2b specified in Sect. 1). Furthermore, the two selected catchments are different in size, signatures (Table 2) and complexity of their hydrological response, so that the influence of the catchment or data properties can be assessed to some degree. To limit the scope of the study, we constrained the analysis to two catchments.

Table 2. Properties of the two case study catchments. P is the precipitation and R_C the runoff coefficient (calculated from cumulative streamflow and precipitation). $Q_{\text{obs, max}}$, $Q_{\text{obs, min}}$ and $\overline{Q}_{\text{obs}}$ are the minimum, the maximum and the average streamflow, respectively. $I_{\text{F, obs}}$ is the flashiness index.

Catchment	Area (km^2)	P (mm a^{-1})	R_C (–)	$Q_{\text{obs, max}}$ (mm h^{-1})	$Q_{\text{obs, min}}$ (mm h^{-1})	$\overline{Q}_{\text{obs}}$ (mm h^{-1})	$I_{\text{F, obs}}$ (–)
Murg	80	1369	0.57	2.7	1×10^{-2}	0.089	0.053
Maimai	0.07	2349	0.62	8.5	1×10^{-4}	0.17	0.13

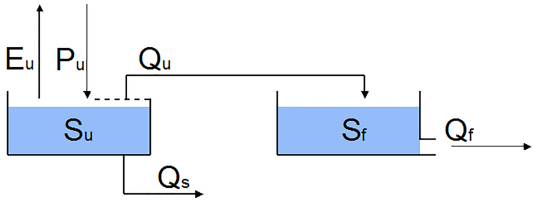

Figure 2. Structure of the deterministic hydrological model used in this study. P_u is the precipitation and E_u the evapotranspiration. S_u represents the active water content of the unsaturated zone, while S_f is a non-linear reservoir representing the fast flow component.

3.2 Deterministic hydrological model

The hydrological model used throughout this study is a simple, lumped bucket model with two reservoirs (Fig. 2), which are meant to represent the unsaturated soil zone and the subsurface flow being fed by it. A slower flow component is included though a linear outflow from the unsaturated zone reservoir directly. Due to its simplicity, and due to the fact that it is not clear whether the chosen model structure is suited for the studied catchment a priori, we expect systemic difficulties in reproducing the observed streamflow dynamics. This is a very common situation in hydrological modelling and it will lead to correlated and potentially heteroscedastic and non-normal errors. This allows us, in principle, to test the error models (Sect. 2.2) under realistic conditions. The streamflow simulated by this deterministic model is denoted as $Q_{\text{det}}(t, \boldsymbol{\theta}) = Q_s(t, \boldsymbol{\theta}) + Q_f(t, \boldsymbol{\theta})$, where Q_s is the slow response of the model, Q_f is the fast response and $\boldsymbol{\theta} = (C_e, S_{\max}, k_u, k_f)$ are the calibrated hydrological parameters. The fluxes (E_u, P_u, Q_u, Q_s, Q_f) and states (S_u, S_f) of the model are given by the following:

$$\begin{aligned}
\frac{\mathrm{d}S_u}{\mathrm{d}t} &= P_u - E_u - Q_u - Q_s, \\
E_u &= C_e E_p \frac{\frac{S_u}{S_{\max}}(1+m)}{\frac{S_u}{S_{\max}} + m}, \\
Q_u &= P_u \left(\frac{S_u}{S_{\max}}\right)^{\beta}, \\
Q_s &= k_u S_u,
\end{aligned} \tag{18}$$

$$\begin{aligned}
\frac{\mathrm{d}S_f}{\mathrm{d}t} &= Q_u - Q_f, \\
Q_f &= k_f S_f^{\alpha},
\end{aligned} \tag{19}$$

where E_p is the potential evapotranspiration. While C_e, $S_{\max}$, k_u and k_f were inferred, m, β and α were kept fixed at 0.01, 3 and 2, respectively. m can be seen as a smoothing parameter and $m = 0.01$ translates to $E_u \approx C_e E_p$ as long as $S_u/S_{\max} \gg 0.01$. $\beta = 3$ and $\alpha = 2$ were found to lead to reasonable results in both investigated catchments and were fixed due to potential interactions with $S_{\max}$ and k_f. The hydrological model was implemented in SUPERFLEX (Fenicia et al., 2011; Kavetski and Fenicia, 2011), a flexible framework for conceptual hydrological models which uses efficient numerical integration schemes.

3.3 Priors

The prior distribution of the parameters was assumed to be composed of independent normal or log-normal distributions with relatively large standard deviations (see Table 3). A unimodal distribution is the more accurate representation of our prior belief than, for example, a uniform distribution over a predefined range, since we do assume that values in the middle of the suspected range are more probable than at its edge. Note that this is primarily a conceptual difference, as large standard deviations were chosen to minimise the influence of the priors on the results.

4 Results

After providing some general results, this section contains a more detailed summary of the results for each of the tested error models. The complete analysis included additional error models and performance metrics, which are included in Appendix B. The supplementary material contains further information on the resulting posterior density estimates of the parameters and Kullback–Leibler divergences of the marginal posterior and prior parameter density estimates.

Figure 3 gives an overview of the difference in flashiness index, the reliability and the relative spread in the calibration and the validation periods for both catchments, all temporal resolutions of the data and all tested error models. Figure 4 provides additional information about the relative

Table 3. Prior distributions of the hydrological and error model parameters applied in all the cases where the respective parameter was used. N = Gaussian normal; LN = log-normal. Where lower and upper boundaries are listed, the distribution is truncated at those values.

Parameter	Distribution	Unit	μ	σ	Lower boundary	Upper boundary
C_e	N	–	1	0.2	0.2	3
S_{max}	LN	mm	148	1086	2.7	1086
k_u	LN	h^{-1}	1.8×10^{-2}	0.13	2.3×10^{-6}	5×10^{-2}
k_f	LN	h^{-1}	0.37	2.7	2.3×10^{-6}	0.37
a	LN	–	0.2	0.2	–	–
b	LN	–	0.1	0.1	1×10^{-2}	0.5
τ_{max}	LN	h	148	1086	0	2000
γ	LN	–	1	0.2	0.1	5
d_f	LN	–	14	17	3	–

error in cumulative streamflow, Δ_Q, and about the Nash–Sutcliffe efficiency, $\widehat{E}_{N,\,det}$. The temporal resolution of the data has a pronounced effect on all the analysed performance metrics. The spread over all the combinations of error models and catchments is larger for higher temporal resolutions (Figs. 3 and 4). Furthermore, the average of each metric indicates decreasing performance for increasing temporal resolution. This loss in performance is more pronounced in the Murg catchment and for error models E2 and E3a than in the Maimai catchment and for other error models. The difference between the two catchments is most clearly visible in $\widehat{E}_{N,\,det}$ (Fig. 4): for 6-hourly and daily resolution of the data, the worst-performing error model in the Maimai catchment has a better $\widehat{E}_{N,\,det}$ than the best-performing error model in the Murg catchment.

4.1 Individual error models

4.1.1 Model E1

E1 tends to strongly overestimate the true flashiness in the case of high temporal resolutions in both catchments (Fig. 3a, b; the difference between the observed and the median of the predicted flashiness index is around -0.4 for both catchments). In terms of reliability, E1 is never the single best of the error models but is always among the best, and it is robust in light of varying temporal resolution (Ξ_{reli} is larger or equal to 0.8 in all the cases; Fig. 3c, d). E1 is also among the error models that provide the least uncertain predictions (average relative spread of 0.41, Fig. 3e, f) and have the smallest Δ_Q (usually between 0 % and -10 %) and the highest $\widehat{E}_{N,\,det}$ overall (Fig. 4). Except for the flashiness index, its performance stays stable for high-frequency data in both catchments. However, the high flashiness index of this model demonstrates the strong violation in the description of the output behaviour despite its good performance regarding the other performance metrics.

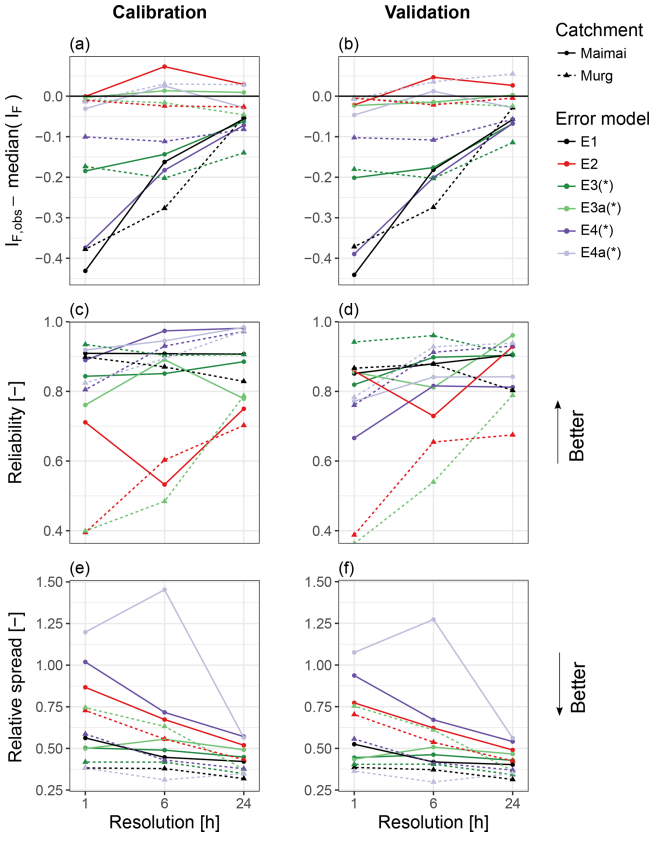

Figure 3. Performance of the error models with respect to the flashiness index, reliability and relative spread for both catchments and all temporal resolutions. P_{err} was smoothed (*) exclusively for hourly data in the Maimai catchment.

4.1.2 Model E2

With the constant correlation assumption made in E2, $I_{F,\,obs}$ is generally well reproduced by I_F, with deviances ranging from -0.03 to 0.07 (Fig. 3a, b). For E2, $\widehat{I}_{F,\,det}$ is often similar to I_F for all temporal resolutions (Tables B1 and B2), indicating that the large part of the flashiness of the model

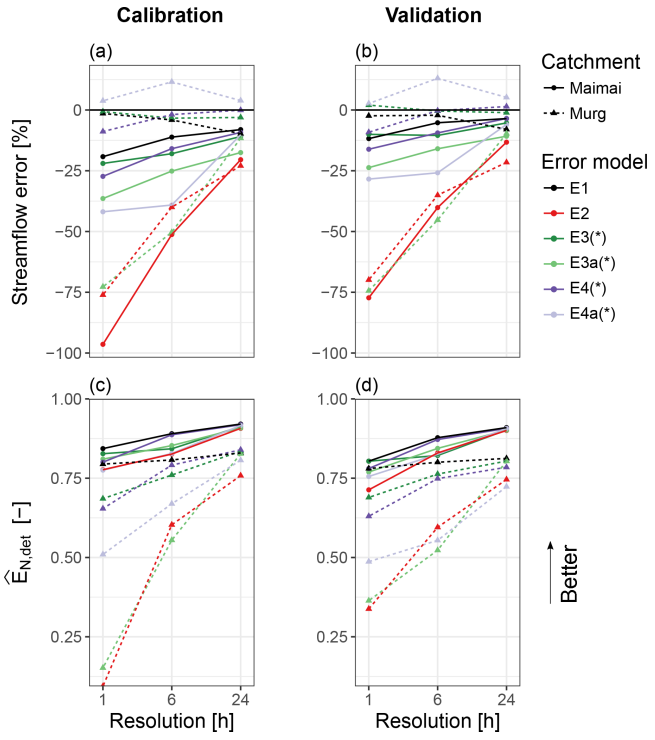

Figure 4. Performance of the error models in terms of the relative cumulative error in streamflow, Δ_Q, and the Nash–Sutcliffe efficiency, $\widehat{E}_{\mathrm{N,\,det}}$, for both catchments and all temporal resolutions. P_{err} was smoothed (*) exclusively for hourly data in the Maimai catchment.

output is due to the hydrological model response and only a small part is due to the stochastic variability added through the error model. Regarding all the other performance metrics, however, E2 is often among the worst-performing error models. For example, in more than half of all the investigated combinations of catchments and temporal resolutions, E2 is the error model with the worst reliability (Fig. 3c, d). E2 has an average relative spread of 0.61 over all the cases, while that of E1 is 0.41. It tends to produce large errors in cumulative streamflow, especially in the case of hourly resolution ($\Delta_Q < -75\,\%$, Fig. 4a, b). The degradation of the streamflow error and $\widehat{E}_{\mathrm{N,\,det}}$ with increasing measurement frequency is very pronounced for E2 compared to the other error models (Fig. 4a–d).

4.1.3 Model E3

E3 generally overestimates the true flashiness; i.e. I_{F} is often larger than $I_{\mathrm{F,\,obs}}$. The difference is around 0.2 for hourly and 6-hourly resolution and a bit less for daily resolution (Fig. 3a, b). The overestimation of the flashiness by E3 is less severe than with E1. E3 results in stable reliability metrics for all temporal resolutions in both catchments: Ξ_{reli} is larger than 0.8 in every case and larger than 0.9 in more than half of the cases (Fig. 3c, d). In the validation period in the Murg catchment, it is the most reliable error model of all (Fig. 3d). The relative spread of E3 is in the range of [0.34, 0.5] in all instances with an average value of 0.43, and it is unaffected by the temporal resolution (Fig. 3e, f). The absolute value of Δ_Q is never larger than 25 % and usually smaller than 10 % (Fig. 4a, b). In terms of $\widehat{E}_{\mathrm{N,\,det}}$, E3 reaches values larger than 0.75 in all cases except for hourly resolution in the Murg catchment, where it is 0.69. All the metrics show stable performance of E3 under increasing measurement frequency (Figs. 3 and 4).

4.1.4 Model E3a

When inferring τ_{min} with error model E3a, we get close correspondence of I_{F} and $I_{\mathrm{F,\,obs}}$ in all cases (Fig. 3a, b; the deviation is never larger than 0.05). In the Maimai catchment, the reliability measure shows stable performance, with values between 0.81 and 0.96 in the validation period (Fig. 3c, d), showing no clear signs of worse performance for high-frequency data. The inferred values of τ_{min} were of the order of 1 d and therefore clearly smaller than τ_{max} (Fig. 7). Furthermore, τ_{min} was consistent among the different temporal resolutions.

In the Murg catchment, on the other hand, we see a degenerating performance of E3a with increasing measurement frequency, with values of $\Xi_{\mathrm{reli}} < 0.5$ for 6-hourly and hourly data (Fig. 3c, d), indicating poor performance. All the other metrics show a similar pattern. The inferred τ_{min} were between 50 and 100 h, where values on the upper end of that range coincided with bad reliability (Fig. 7).

4.1.5 Model E4

The stochastic model realisations with E4 tend to overestimate the true flashiness index; the difference between $I_{\mathrm{F,\,obs}}$ and I_{F} is usually between -0.2 and -0.1 (Fig. 3a, b). I_{F} is often much larger than $\widehat{I}_{\mathrm{F,\,det}}$ in the Murg catchment (Table B1), indicating that a relatively large part of the flashiness is accounted for by the error model and less by the hydrological model in that case. This manifests in smaller values of $\widehat{E}_{\mathrm{N,\,det}}$ with E4 compared to E1 (e.g. 0.65 for E4 with hourly resolution compared to 0.79 with E1, Fig. 4c). In the Maimai catchment, the hydrological model captures a larger part of the variability than in the Murg catchment, and the difference between I_{F} and $\widehat{I}_{\mathrm{F,\,det}}$ is smaller (Table B2). Concerning the reliability, Ξ_{reli} is generally larger than 0.8, indicating well-conditioned predictive distributions, except in the validation period for hourly resolution (Fig. 3c, d). In the Maimai catchment, reliability is better in the calibration period than in the validation period, which is a sign of over-fitting. Especially for daily resolution, E4 provides very good reliability in the calibration period in both catchments ($\Xi_{\mathrm{reli}} > 0.97$, Fig. 3c). The average relative spread of E4 is 0.60. Δ_Q is not more extreme than $-27\,\%$ in any case and usually less severe than

20 % (Fig. 4a, b). A slight degradation of Δ_Q with increasing frequency of the data can be observed.

4.1.6 Model E4a

E4a results in I_F that are very close to the observed flashiness in all cases: the difference is never more extreme than 0.05 (Fig. 3a, b). $\widehat{I}_{F,\,det}$ is often smaller than $I_{F,\,obs}$ in the Murg catchment, which, similar to in E4, is an indication that most of the variability is explained by the error model and not the hydrological model. Ξ_{reli} is always larger than 0.8 (Fig. 3c, d) except for the validation period with hourly resolution in both catchments (Fig. 3d). Similar to E4, we can see a tendency for over-fitting with E4a in the Maimai catchment: in the calibration period, reliability values of 0.98, 0.95 and 0.92 are reached, while the validation results in values of 0.84, 0.84 and 0.77 for daily, 6-hourly and hourly resolutions, respectively (Table B2). A look at the relative spread (Fig. 3e, f) shows that E4a leads to unrealistically large prediction uncertainty in the Maimai catchment for 6-hourly and hourly resolution but that it is among the most precise error models in the Murg catchment. Similarly, E4a produces relatively large errors in cumulative streamflow in the Maimai catchment, but very small ones in the Murg catchment (Fig. 4a, b). Opposed to that, $\widehat{E}_{N,\,det}$ is better than 0.75 in all cases in the Maimai catchment, while it reaches values as low as 0.5 for hourly resolution in the Murg catchment (Fig. 4c, d).

4.2 Relaxing the constant-correlation assumption

Error model E3, which accounts for reduced correlation of errors during the precipitation events, leads to an overall improvement in the investigated performance metrics (except I_F) compared to E2, which assumes constant correlation (Figs. 3 and 4). For example, the reliability for hourly resolution in the Murg catchment is 0.94 and 0.39 for E3 and E2, respectively (Fig. 3c, d). In contrast to E2, the performance of E3 does not show systematically worse performance for high-frequency data. In fact, E3 and E1 show a similar stability in performance, but E3 provides more realistic estimates of the correlation during recessions and baseflow, leading to a better estimate of I_F (Fig. 3a, b). Figure 6 shows typical results of E2 and E3 with respect to streamflow bias, visible as a bias in η (Fig. 6a, b), and posterior correlation between heteroscedasticity and correlation parameters a and τ_{max} (Fig. 6c, d). Note also the smaller standard deviation (parameter a) resulting from E3 (Fig. 6d). Additional results about the standardised innovations of η are available in the Supplement.

Figure 5 compares the predicted hydrographs of E1, E2 and E3a in the Maimai catchment using hourly data. In this case, allowing for different characteristic correlation times during precipitation events and dry periods (E3a, Fig. 5c) leads to better-behaved error bands compared to the constant correlation assumption (Fig. 5b) and to more realistic stochastic output of the model than with the zero-correlation assumption (Fig. 5c). Note that E3a results in better estimates of I_F than E3, since it considers correlation during precipitation events ($\tau_{min} > 0$).

In the Murg Catchment, inferring τ_{min} resulted in a degenerative performance for high-frequency data, which were also linked to higher values of τ_{min} (Fig. 7). The posterior estimates of τ_{max} depend on the resolution in both catchments. While large τ_{min} coincides with the worst reliability, large τ_{max} was also obtained together with good reliability (Fig. 7). The effect of τ_{min} on the relative cumulative streamflow error is shown in Fig. 8 for 6-hourly data in the Murg catchment. The streamflow error starts to increase for $\tau_{min} > 10\,h$ and at the same time $\widehat{E}_{N,\,det}$ decreases (not shown), approaching that of E2.

4.3 Relaxing the assumption of normality

Relaxing the assumption of normality by inferring γ and d_f (E4 and E4a) had a mixed effect on the numeric performance indices analysed in this study. When $\tau_{min} = 0$, including skewness and kurtosis (E4) often led to a better reliability in the calibration period (Fig. 3c), but a worse reliability in the validation period (Fig. 3d) compared to the assumption of a normal distribution with E3. Predictions with E4 generally had a smaller spread than those with E3; e.g. Ω_{spread} was around 0.5 with E3 and 1.0 with E4 for hourly resolution in the Maimai catchment (Fig. 3e, f). When τ_{min} was inferred additionally, the non-normal case (E4a) showed better performance metrics than the normal case (E3a) in the Murg catchment, but worse ones in the Maimai catchment. E4 and E4a in the Maimai catchment were the only cases that showed a pronounced difference between calibration and validation, which is a sign of overfitting. A visual inspection of the QQ plots of η revealed that E4 and E4a successfully reduced some very heavy outliers that strongly violated the assumption of normality. In both catchments, the inferred γ were in the range of [1.5, 2.8] for E4 and E4a. The values at the upper end of this spectrum were reached for hourly resolutions, and they were associated with underestimation of the peak flows by the deterministic hydrological model, reflected in reduced $\widehat{E}_{N,\,det}$. For example, E4a resulted in $\gamma \approx 2.5$, $\widehat{E}_{N,\,det} = 0.5$ and an underestimation of peak flows by the hydrological model for hourly data in the Murg catchment. Inferred d_f were always at or close to the lower limit of 3, which is indicative of heavy outliers.

Regarding the location of D_Q with respect to Q_{det}, the assumption in Eq. (9a) led to better results than Eq. (9b) in the Murg catchment. For example, Ξ_{reli} with E4a is 0.22 or 0.87 when applying Eq. (9a) or (9b), respectively (Table B1). In the Maimai catchment, the opposite is true: Ξ_{reli} is 0.32 or 0.23 with Eq. (9a) or (9b), respectively (Table B2). The difference between results obtained with Eqs. (9a) and (9b) is generally larger for higher frequency of the data.

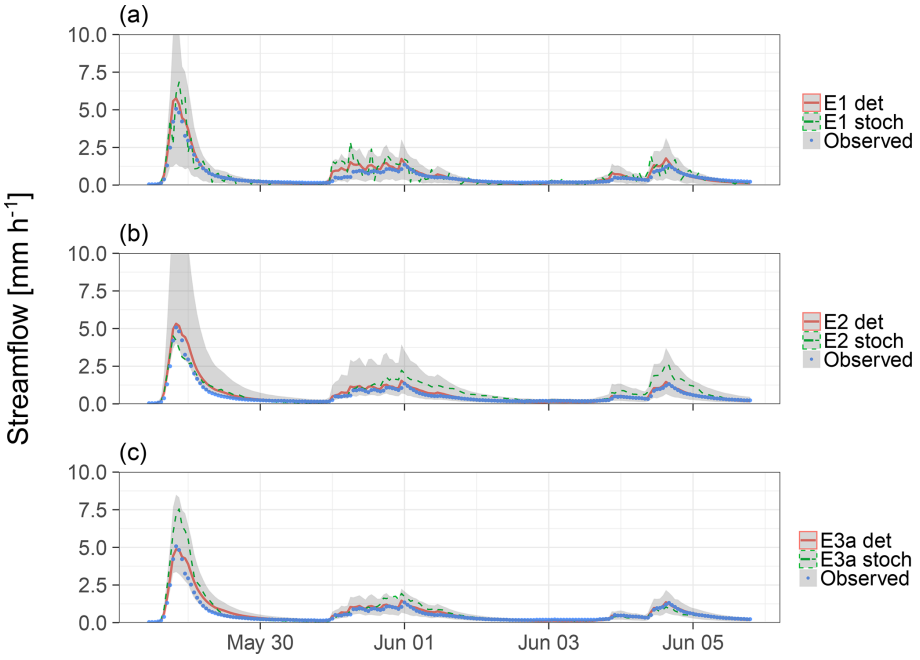

Figure 5. Streamflow predictions with hourly resolution in the Maimai catchment in a part of the validation period (1993) obtained with error models E1 (a), E2 (b) and E3a (c). Deterministic predictions with the parameter values at the maximum posterior density are shown together with the 90 %-confidence bands and one single stochastic streamflow realisation for each of the error models.

5 Discussion

5.1 Presence and absence of autocorrelation

Assumptions about the presence (E2) and absence (E1) of autocorrelation in η were shown to have profound effects on the quality of the prediction in the cases investigated in this study. Neglecting autocorrelation leads to close correspondence between $\widetilde{Q}_{\mathrm{det}}$ and Q_{obs} in terms of the Nash–Sutcliffe efficiency and to relatively well-fulfilled assumptions about the distribution of η in the uniform space (i.e. small values of Ξ_{reli}). However, major assumptions of the underlying statistical model are clearly violated. Most striking is the violation of the zero correlation assumption (Fig. 9b), which translates into unrealistic fluctuations of the stochastic streamflow predictions (Fig. 5a). Note that E1 also comes with disadvantages related to operational forecasts, where one can make more accurate predictions for streamflow in the near future given an error in previous streamflows when accounting for correlated errors (Del Giudice et al., 2013). This effect was not analysed in this study.

Accounting for the fact that η is obviously autocorrelated, and therefore describing it by a Gaussian process with constant autocorrelation (E2), comes with additional difficulties. These include a strong interaction of the hydrological water balance parameter, C_{E}, with autocorrelation, τ_{max}. In addition, we observed a strong posterior correlation between the parameter for autocorrelation, τ_{max}, and heteroscedasticity, a (Fig. 6c). This correlation in the posterior parameter distribution coincided with systematic overprediction of streamflow. E2 also showed smaller E_{N} and $\widehat{E}_{\mathrm{N,det}}$, and worse Δ_Q compared to E1 (Fig. 4). Evin et al. (2013), who tested an error model similar to E2 on daily data, obtained very similar results in terms of interactions of water balance parameters with correlation and heteroscedasticity parameters. The reasons for those problems are still poorly understood. Failing to reproduce the problems under synthetic conditions, Evin et al. (2014) suggest that the "nonrobustness of the joint approach" might be caused by "structural errors in the hydrological and/or error models". Based on case studies with daily data, they find that (i) the catchments where these problems are absent are all wet catchments with relatively high runoff coefficients and low ephemerality. To this, we can add that (ii) the performance of the corresponding error model in our study (E2) strongly degrades for higher data frequency within two relatively wet catchments.

5.2 Non-stationarity of autocorrelation

Figure 9 visualises one potential reason for the degrading performance of E2 for high-frequency data: our assumptions about the stochastic process (OU process with constant correlation time τ) seem to be much better fulfilled for the daily (Fig. 9a) than for the hourly (Fig. 9b) data. In the latter case, a visual assessment of $\eta(t)$ obtained with E1 reveals strongly reduced auto-correlation during storms compared to inter-storm periods. Yang et al. (2007) made similar observations. This raises the hypothesis that the neglect of

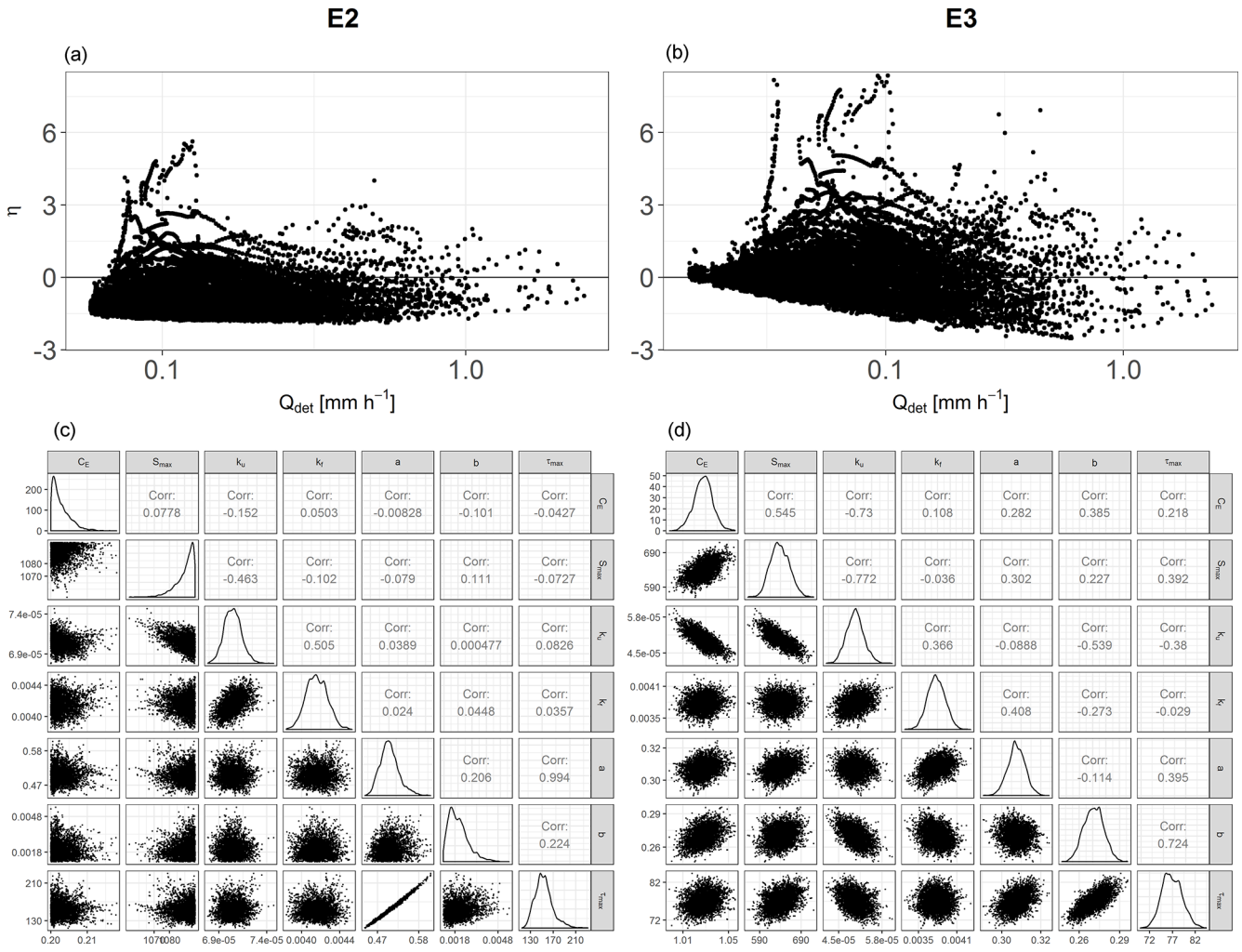

Figure 6. Transformed residuals, η, as a function of modelled streamflow **(a, b)** and correlation structure of the posterior parameter sample **(c, d)** resulting with error models E2 **(a, c)** and E3 **(b, d)** for data with hourly resolution in the Murg catchment.

non-stationarity of the autocorrelation is a major deficit of conventional error models, which leads to the previously encountered problems in the joint inference of autoregressive and hydrological model parameters mentioned in Sect. 5.1.

What is the physical explanation for non-stationary autocorrelation of the errors η? The autocorrelation of errors in streamflow is primarily caused by the memory effect of errors in storage (Kavetski et al., 2003). Since this memory effect of a catchment during precipitation events can be expected to be different from that during dry weather, the correlation of the errors in streamflow can be expected to be different as well. The degree of change of the correlation may depend on multiple factors, like the hydrological model used, the precipitation intensity or volume, the extent to which the precipitation signal is filtered by the catchment, time lags between precipitation and runoff, and potentially other factors. Most probably, the mentioned factors will lead to smaller correlation during wet periods and larger ones during dry periods.

A very simple way of considering this reduced correlation (E3) provides strongly improved results compared to the assumption of stationary correlation (Sect. 4.2). This indicates that neglect of the non-stationarity of the autoregressive parameter is a substantial shortcoming of conventional error models, which causes, at least partly, the well-known problems related to joint inference. Note that non-stationary correlation can also be implemented in other existing likelihood functions and does in principle not require the use of the proposed theoretical framework described in Sect. 2.1.

To challenge this hypothesis, one could argue that the improved performance of E3 (compared to E2) might also be achieved when reducing τ during completely arbitrary time intervals instead of precipitation events. This would dismiss the hypotheses that the precipitation has a direct influence on τ and that considering this influence leads to a better in-

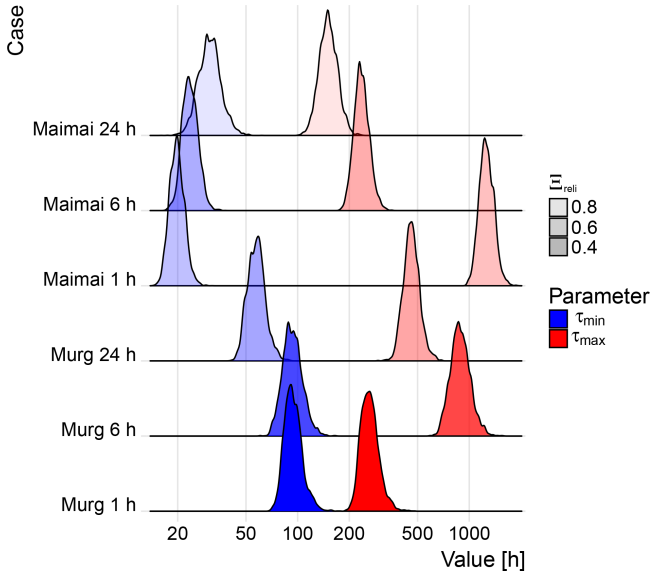

Figure 7. Marginal posterior densities of $\tau_{\min}$ and $\tau_{\max}$, and corresponding reliability measures Ξ_{reli} in the validation period resulting from error model E3a in all combinations of catchments and temporal resolutions.

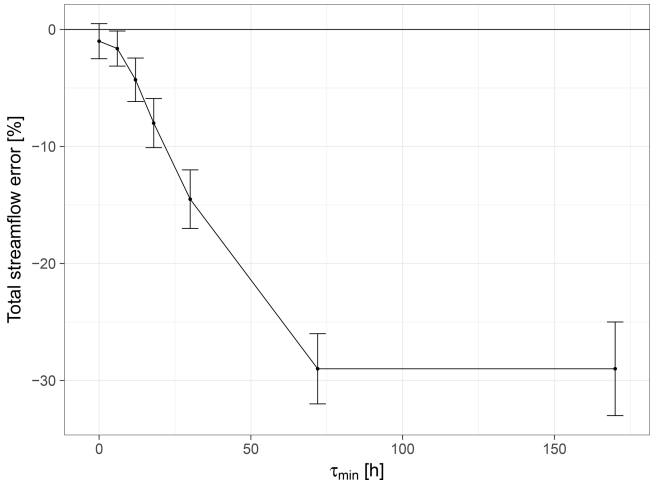

Figure 8. Relationship between the fixed correlation time during precipitation events, $\tau_{\min}$, and the total streamflow error, Δ_Q, for 6-hourly data resolution in the Murg catchment. Each point corresponds to a full inference and prediction procedure. The error bars span two standard deviations of 500 stochastic predictions. E3 corresponds to $\tau_{\min} = 0$ and E2 to $\tau_{\min} = \tau_{\max} \approx 170$ h.

ference behaviour. To test this, we shifted P_{err} (Eq. 11) substantially in time, so that it would not correspond to the observed precipitation P anymore, while still keeping the major properties (duration and intermittency) of the time intervals during which τ is reduced. Then, inference was performed with E3 again. The low Nash–Sutcliffe efficiency and the high streamflow error of the stochastic predictions in that case (E3† in Table B2) shows that it is indeed important to reduce τ during the precipitation events and not during arbitrary periods with the same intermittency and duration as the precipitation events. With the shifted P_{err}, the resulting $\tau_{\max}$ (≈ 145 h) was much smaller than the original $\tau_{\max}$ (≈ 1400 h), confirming the hypothesis of reduced correlation time of errors in streamflow during precipitation events.

One could also argue that the improved performance of E3 compared to E2 is primarily due to assuming reduced autocorrelation during periods with strong outliers (i.e. storm events) and that those outliers (visible in Fig. 6) should be accounted for by appropriate values of γ and d_{f}, instead of reducing their influence by neglecting correlation in the periods they appear. Or, similarly said, if the autoregressive process with constant correlation is applied to appropriately standardised residuals, which are marginally normally distributed, it should not cause any problems. To explore this possibility, we performed some experimental analysis for hourly resolution in the Murg catchment: we modified E1 by fixing $\gamma = 1.5$ and $d_{\text{f}} = 5$ (E1^{+}). This led to a well-conditioned η and performance metrics that were comparable to or better than those of E1 (Table B1). Then, we inferred τ under the assumption of constant correlation, while skewness and kurtosis were kept fixed at the values given above (E2^{+}). The resulting performance metrics and a visual assessment of the hydrographs revealed strong deficiencies in this approach compared to E3 and to E1^{+} (Table B1). This indicates that it is not enough to ensure that the marginal distributions of errors is sufficiently well captured before applying an autoregressive process, but that it is also important to account for a potential non-stationarity of the correlation of the errors. Note that the distributional parameters of D_Q (e.g. γ or d_{f}) could also be non-stationary (Wani et al., 2019).

It is still unclear what the optimal parameterisation of a time-dependent correlation could be. Using the input to directly inform the correlation structure of the output requires knowledge of how the catchment transforms the signal. For example, there could be a significant time lag between precipitation and streamflow, which would have to be taken into account in Eq. (11). For the Maimai catchment, we found that using a smoothed version of P_{err} in Eq. (11) improved the performance of error models E3 and E4 in the case of hourly resolved data (Table B2). For the coarser resolutions in the Maimai catchment, and for all the tested resolutions in the Murg river, transforming P_{err} in a similar way did not lead to a remarkable change in the results. The influence of possible transformations of P_{err} to account for the filtering effect of the catchment was not systematically investigated in this study.

5.3 Inference of $\tau_{\min}$

The fact that $\tau_{\min}$ (Eq. 11) could only be inferred with partial success shows that there are still problematic interactions among parameters controlling the correlation of the errors and hydrological model parameters. Figure 7 indicates that

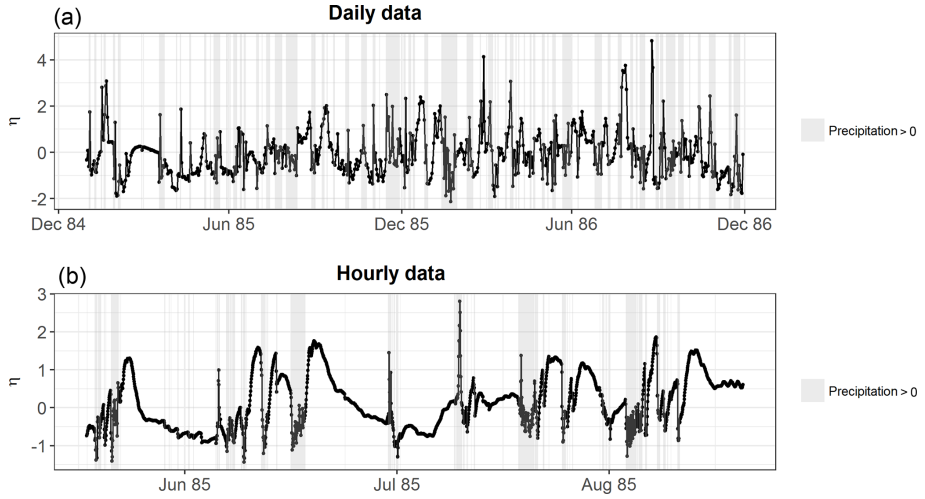

Figure 9. Time series of η corresponding to the parameter values at the maximum posterior density obtained with E1 in the Maimai catchment for daily and hourly resolution. Intervals where $P > 0$ are shaded in grey.

those problems are more related to $\tau_{\min}$ than to $\tau_{\max}$, since higher values of $\tau_{\min}$ tend to coincide with bad performance. Or, in more general terms, the previously encountered problems in the joint inference of hydrological and correlation parameters (Evin et al., 2013) seem to originate from precipitation periods, not from dry periods. The fact that the inference of $\tau_{\min}$ is more successful in the Maimai catchment (Sect. 4.1.4), which has the simpler hydrological response, suggests that the realism of a hydrological model facilitates the successful inference of the correlation parameters.

These findings call for additional investigations into the issue of non-stationary correlation, potentially exploring other relationships between τ and P or Q_{det}. Making τ dependent on Q_{det} instead of P would have the advantage that potential low-pass filtering or time lag between precipitation and streamflow are taken care of by the hydrological model and need not be considered anymore in the error model. We performed some exploratory analysis in that direction, so far with limited success.

5.4 Shape of the distribution D_Q

Relaxing the assumption of marginal normality of Q_{obs} given Q_{det} successfully reduced some very heavy outliers that strongly violated that assumption. However, this did not always translate to improved distributional assumptions in the uniform space, where Ξ_{reli} is calculated. We suspect that the presence of strong outliers (large η) under the normal assumption led to the strong right-skew of D_Q when inferring γ and d_{f}, which was less appropriate for the rest of the distribution of observed streamflows. In that case, a different distributional shape for D_Q would be more appropriate, e.g. a mixture distribution, which allows for some heavy tails on the upper side without skewing the central body too much to the right. Testing other distributional shapes for D_Q was beyond the scope of this study, however. Note that heavy outliers (i.e. $\eta \gg 0$) do not necessarily correspond to high streamflow; in both catchments the largest η were observed during medium to low flows (Fig. 6a, b), namely during small peaks of observed streamflow that were not captured by the model.

The ranking in performance of the two options to either place the mean or the mode of D_Q at Q_{det} (Eq. 9) was different for the two analysed catchments. The former led to better results in the Murg catchment, while the latter seemed preferable in the Maimai catchment. Ideally, we would like to satisfy both conditions, but this is obviously not possible when D_Q is skewed.

Regarding the choice of the type of the distribution D_Q, recall that $Q(t) \sim D_Q(Q_{\text{det}}(t), \boldsymbol{\psi})$. A distribution type with positive support would be a desirable alternative to the skewed Student's t distribution, since it would ensure positive streamflow without the need to assign the probability of $Q < 0$ to $Q = 0$. If, additionally, $E[Q(t)] = Q_{\text{det}}(t)$, mass conservation would be guaranteed (since the applied hydrological model conserves mass). Some limited exploration in this direction with a lognormal distribution lead to unsatisfactory fits (results not shown). This might be due to the unrealistically strong right-skew needed to account for cases where $Q_{\text{obs}}(t) \gg Q_{\text{det}}$ when using a distribution with positive support and mean equal to Q_{det}. Thus, in our experience, the non-negativity of streamflow observations (for non-tidal rivers) makes the conservation of mass difficult at very low modelled streamflow if there is a considerable observation error.

6 Conclusions

We presented and evaluated a flexible framework for probabilistic model formulations (i.e. likelihood functions) to describe the total uncertainty of the output of deterministic hydrological models. This framework allows us to consider heteroscedastic errors with non-stationary correlation, non-equidistant observations and zero probability for negative streamflow. It does so by allowing for arbitrary and explicit marginal distributions for the observed streamflow at each point in time. For experts, it is easier to parameterise these marginal streamflow distributions than the distribution characterising the autoregressive model or some non-intuitive transformations like the Box–Cox transformation. The consistent implementation of this framework was successfully checked with a synthetic case study.

Using a simple deterministic hydrological bucket model and two case study catchments, the flexible framework was used to systematically test different error models on real-world data. Those error models represented various assumptions about the statistical properties of the errors in terms of autocorrelation, skewness and kurtosis. The assumptions were found to have a profound effect on the quality of the predictions. The key findings are as follows:

1. We confirmed that, as shown in previous work by various authors, accounting for autocorrelation with conventional approaches (represented by model E2) can lead to worse predictions than omitting autocorrelation (model E1). For example, model E2 had errors in cumulative streamflow of 76 % in the Murg catchment and 96 % in the Maimai catchment for hourly resolution in the calibration period. With model E1, in comparison, those errors were 1 % and 19 %, respectively. However, this result is unsatisfactory as there is clearly visible autocorrelation in the residuals that invalidates the model E1.

2. We showed that the predictions of conventional approaches to deal with autocorrelation worsen significantly as the temporal resolution increases. For example, the performance of model E2 in terms of the Nash–Sutcliffe efficiency decreases from 0.76 to 0.09 in the calibration period when moving from daily to hourly data resolution. In comparison, the performance of model E1 remains relatively stable (Nash–Sutcliffe efficiency decreases from 0.83 to 0.79).

3. Since rapid changes in a catchment's storage change its memory, errors in streamflow are expected to show different correlations during precipitation events and dry weather. Based on the hypothesis that this non-stationarity increases when going from daily to hourly resolution, neglecting non-stationarity of correlation is the likely cause for finding 2.

4. Accounting for non-stationarity in autocorrelation significantly alleviated the observed problems of finding 2. In particular, allowing for the autocorrelation to be lower during wet than during dry periods (models E3 and E4) led to more stable behaviour across time resolutions. For example, volume errors for model E3 in the Murg catchment were not larger that 3 % for all three investigated temporal resolutions. However, inferring the characteristic correlation time during precipitation events (model E3a) provided good results in only one of the two investigated catchments. Keeping that correlation fixed (model E3) could be seen as a pragmatic option with stable performance.

5. If the problems mentioned in finding 1 can be avoided, accounting for autocorrelation results in more realistic characteristics of model output than omitting autocorrelation, which is confirming previous work. In particular, signatures such as the flashiness index are much better represented when including autocorrelation. For example, for an observed value of the flashiness index of 0.13 in the Maimai catchment in the calibration period, model E3a provided a value of 0.13, whereas model E1 resulted in a much larger value of 0.56.

6. Inferring the skewness and kurtosis of a skewed Student's t distribution can lead to better-fulfilled distributional assumptions about the errors. In our case study, this expectation was partly fulfilled for daily data, but not for data of higher frequency. For hourly data, for example, more freedom with respect to the shape of the distribution actually lead to less accurate representation of the observed distribution.

These results contribute to a better characterisation of the residual errors of deterministic hydrological models. However, some questions remain. It still has to be shown to what degree the findings of this study are generalisable to a larger and more diverse set of catchments and to different hydrological models. A comparison of the presented approach to existing frameworks based on different assumptions, like the generalized likelihood framework, would yield further insights. Furthermore, it is still unclear how the non-stationary autocorrelation should ideally be parameterised. The chosen approach, where we alternate between two values of the autoregressive parameter based on whether there is precipitation or not, might lead to problems in catchments with strong lags between precipitation and streamflow. In those cases, defining the autoregressive parameter as a function of modelled streamflow might be more suitable. Furthermore, future studies could investigate different approaches to describe non-stationary correlation or distributions other than the Gaussian and the skewed Student's t. Overall, this study confirms previously encountered difficulties in finding a parameterisation of an additive error term that adequately describes the effects of intrinsic stochasticity.

Data availability. The data of the Maimai catchment can be obtained from Jeffrey McDonnell (Associate Director at Global Institute for Water Security and Professor at the School of Environment and Sustainability at the University of Saskatchewan, https://www.usask.ca/watershed/index.php, last access: 26 April 2019) upon request. The meteorological data of the Murg catchment can be obtained through MeteoSwiss, the Swiss Federal Office of Meteorology and Climatology. The streamflow data of the Murg catchment are available at FOEN, the Swiss Federal Office for the Environment.

Appendix A: Derivation of the likelihood function

To derive the conditional distribution of $Q(t_i)|Q(t_{i-1})$ (and construct the likelihood function by iteratively multiplying the conditional probability densities), we have to propagate the distribution $\eta(t_i)|\eta(t_{i-1})$ given by Eq. (4) to the streamflow using the (inverse) transformation η_{trans} given by Eq. (2).

In simplified notation (which makes it easier to get the key idea without getting in notational details), we get the following:

$$f\big(Q(t_i)|Q(t_{i-1})\big) = f\big(\eta(t_i)|\eta(t_{i-1})\big)\frac{\mathrm{d}\eta(t_i)}{\mathrm{d}Q(t_i)} = f_{\mathrm{OU}}\big(\eta(t_i)|\eta(t_{i-1})\big)\frac{f_{D_Q}\big(Q(t_i)\big)}{f_{\mathrm{N}(0,\,1)}\big(\eta(t_i)\big)}, \tag{A1}$$

where, in the final equation, f_{OU} refers to the standard Ornstein–Uhlenbeck process defined by Eq. (4) and the ratio of the densities f_{D_Q} and $f_{\mathrm{N}(0,\,1)}$ results from the derivative and inner derivative of the transformation given by Eq. (2) (the derivative of cumulative distribution functions are the corresponding probability densities).

With explicit notation of functions and arguments, we get

$$\begin{aligned}
&f\big(Q(t_i)\mid Q(t_{i-1}),\boldsymbol{\theta},\boldsymbol{\psi}\big)\\
&= f\Big(\eta_{\text{trans}}\big(Q(t_i),\,Q_{\text{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\big)|\eta_{\text{trans}}\big(Q(t_{i-1}),\,Q_{\text{det}}(t_{i-1},\boldsymbol{\theta}),\boldsymbol{\psi}\big)\Big)\\
&\quad \frac{\mathrm{d}\eta_{\text{trans}}}{\mathrm{d}Q}\big(Q(t_i),\,Q_{\text{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\big)\\
&= f_{\mathrm{N}\left(\eta_{\text{trans}}\left(Q(t_{i-1}),\,Q_{\text{det}}(t_{i-1},\boldsymbol{\theta}),\boldsymbol{\psi}\right)\exp\left(-\frac{t_i-t_{i-1}}{\tau}\right),\,\sqrt{1-\exp\left(-2\frac{t_i-t_{i-1}}{\tau}\right)}\right)}\\
&\quad \Big(\eta_{\text{trans}}\big(Q(t_i),\,Q_{\text{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\big)\Big)\\
&\quad \cdot\frac{f_{D_Q\left(Q_{\text{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\right)}\big(Q(t_i)\big)}{f_{\mathrm{N}(0,\,1)}\Big(\eta_{\text{trans}}\big(Q(t_i),\,Q_{\text{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\big)\Big)}.
\end{aligned} \tag{A2}$$

This corresponds to the first sub-equation of Eq. (7). The order of the factors was changed in Eq. (7) to emphasise the product of the marginal distribution f_{D_Q} with a modification factor that tends to unity if $t_i - t_{i-1}$ becomes much larger than τ. The other sub-equations in Eq. (7) consider truncating the streamflow distribution at zero and assigning a point mass corresponding to the integral of the tail below zero to a streamflow of zero.

Appendix B: Complete results

Table B1. Murg: summary of the predictions in the calibration and the validation period made with error models E1–E4 for different temporal resolutions of the hydrological data. Values are medians (and standard deviations) of the quality indices of the deterministic model output for the maximum posterior parameters, as well as those of 500 streamflow realisations produced with the full posterior parameter distributions. Recall that smaller values of Ξ_{reli} and Ω_{spread} indicate better results.

	Calibration								Validation							
Case	Ξ_{reli}	Ω_{spread}	$\widehat{E}_{\text{N,det}}$	E_{N}	$\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	I_{F}	$I_{\text{F, obs}}$	Ξ_{reli}	Ω_{spread}	$\widehat{E}_{\text{N,det}}$	E_{N}	$\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	I_{F}	$I_{\text{F, obs}}$
24 h E1	0.83	0.32	0.83	0.68(0.04)	−10	−10(2.3)	0.29	0.36(0.01)	0.31	0.8	0.31	0.81	0.63(0.04)	−8	−8(2.4)	0.29	0.36(0.01)	0.33
24 h E2	0.7	0.43	0.76	0.48(0.1)	−23	−23(5)	0.29	0.34(0.01)	0.31	0.68	0.42	0.75	0.41(0.1)	−21	−22(5.3)	0.28	0.33(0.01)	0.33
24 h E3	0.91	0.35	0.84	0.65(0.04)	−3	−3(1.6)	0.3	0.45(0.01)	0.31	0.91	0.34	0.8	0.59(0.04)	−1	−1(1.7)	0.29	0.44(0.02)	0.33
24 h E3a	0.79	0.37	0.83	0.62(0.06)	−11	−12(3.5)	0.29	0.36(0.01)	0.31	0.79	0.37	0.8	0.56(0.05)	−8	−10(3.7)	0.29	0.35(0.01)	0.33
24h $\widetilde{\text{E4}}$	0.96	0.35	0.83	0.66(0.1)	5	−1(1.7)	0.27	0.41(0.02)	0.31	0.91	0.34	0.78	0.59(0.16)	6	1(1.9)	0.26	0.4(0.02)	0.33
24 h E4	0.97	0.38	0.84	0.65(0.19)	0	0(1.6)	0.24	0.39(0.02)	0.31	0.93	0.37	0.78	0.57(0.1)	2	1(1.8)	0.23	0.38(0.02)	0.33
24 h $\widetilde{\text{E4}}$a	0.96	0.38	0.76	0.6(0.16)	16	−3(3.7)	0.22	0.29(0.01)	0.31	0.85	0.38	0.64	0.48(0.25)	18	−2(4)	0.2	0.28(0.01)	0.33
24 h E4a	0.97	0.35	0.81	0.66(0.37)	5	4(2.6)	0.2	0.28(0.02)	0.31	0.94	0.35	0.72	0.56(1.4)	6	5(2.8)	0.19	0.27(0.02)	0.33
6 h E1	0.87	0.38	0.81	0.59(0.03)	−4	−4(0.8)	0.12	0.44(0.01)	0.16	0.88	0.37	0.8	0.57(0.02)	−2	−2(0.8)	0.12	0.43(0.01)	0.16
6 h E2	0.6	0.56	0.6	0.13(0.15)	−34	−40(5.7)	0.14	0.18(0)	0.16	0.65	0.54	0.6	0.05(0.14)	−30	−35(5.7)	0.14	0.18(0)	0.16
6 h E3	0.9	0.42	0.76	0.5(0.04)	−3	−3(1.4)	0.15	0.36(0.01)	0.16	0.96	0.41	0.76	0.48(0.03)	0	0(1.5)	0.14	0.36(0.01)	0.16
6 h E3a	0.48	0.63	0.55	−0.03(0.17)	−41	−50(7.3)	0.14	0.18(0)	0.16	0.54	0.61	0.52	−0.18(0.2)	−36	−45(7.5)	0.14	0.18(0)	0.16
6 h $\widetilde{\text{E4}}$	0.95	0.38	0.79	0.62(0.1)	7	−2(1.4)	0.1	0.27(0.01)	0.16	0.9	0.37	0.74	0.56(0.08)	8	0(1.4)	0.1	0.27(0.01)	0.16
6 h E4	0.93	0.43	0.79	0.59(0.1)	−2	−2(1.4)	0.08	0.27(0.01)	0.16	0.91	0.41	0.75	0.52(0.13)	0	0(1.5)	0.08	0.27(0.01)	0.16
6 h $\widetilde{\text{E4}}$a	0.95	0.4	0.63	0.51(0.64)	25	2(3.1)	0.07	0.14(0.01)	0.16	0.92	0.38	0.45	0.35(0.19)	27	5(3.2)	0.06	0.13(0.01)	0.16
6 h E4a	0.89	0.31	0.67	0.58(0.07)	12	12(2.1)	0.06	0.13(0)	0.16	0.93	0.3	0.55	0.45(0.08)	13	13(2.2)	0.05	0.12(0)	0.16
1 h E1	0.9	0.38	0.79	0.54(0.01)	−1	−1(0.4)	0.03	0.43(0)	0.05	0.87	0.39	0.78	0.58(0.02)	−2	−2(0.5)	0.03	0.43(0.01)	0.06
1 h E1$^+$	0.94	0.33	0.75	0.55(0.02)	14	−4(0.4)	0.02	0.33(0)	0.05	0.92	0.33	0.74	0.58(0.02)	12	−6(0.5)	0.02	0.33(0)	0.06
1 h E2	0.39	0.73	0.09	−0.9(0.33)	−61	−76(7.8)	0.04	0.06(0)	0.05	0.39	0.7	0.34	−0.28(0.24)	−56	−70(8.8)	0.04	0.06(0)	0.06
1 h E2$^+$	0.52	0.2	0.5	0.43(0.05)	30	20(1.8)	0.01	0.05(0)	0.05	0.76	0.22	0.59	0.53(0.05)	22	10(2.3)	0.01	0.05(0)	0.06
1 h E3	0.93	0.42	0.69	0.45(0.02)	5	−1(1.3)	0.04	0.22(0)	0.05	0.94	0.4	0.69	0.52(0.02)	7	2(1.6)	0.04	0.24(0.01)	0.06
1 h E3a	0.4	0.74	0.15	−0.7(0.29)	−61	−73(8.8)	0.03	0.06(0)	0.05	0.36	0.75	0.36	−0.31(0.25)	−56	−74(10.7)	0.03	0.06(0)	0.06
1 h E3a*	0.38	0.74	0.21	−0.6(0.26)	−62	−72(8.8)	0.03	0.06(0)	0.05	0.35	0.75	0.39	−0.27(0.22)	−56	−75(9.4)	0.03	0.06(0)	0.06
1 h $\widetilde{\text{E4}}$	0.51	0.96	0.45	−0.29(0.26)	30	−47(4.3)	0.01	0.18(0)	0.05	0.49	0.93	0.48	−0.02(0.23)	28	−45(4.9)	0.01	0.19(0)	0.06
1 h E4	0.8	0.59	0.65	0.41(0.26)	−8	−9(2.5)	0.01	0.15(0)	0.05	0.76	0.56	0.63	0.47(0.07)	−8	−9(2.5)	0.01	0.16(0)	0.06
1 h $\widetilde{\text{E4}}$a	0.15	1.85	0.49	−5.58(0.97)	15	−205(15.1)	0.01	0.07(0)	0.05	0.13	1.78	0.49	−3.8(0.83)	13	−200(17.1)	0.01	0.08(0)	0.06
1 h E4a	0.82	0.38	0.5	0.4(0.03)	4	4(2.1)	0.01	0.06(0)	0.05	0.78	0.36	0.49	0.41(0.03)	4	3(2.2)	0.01	0.06(0)	0.06

* Smoothing $P_{\text{err}}(t)$ with a moving-average window of size 5 h before applying Eq. (11). $\widetilde{\ }$ Denotes the option where $\text{mode}(D_Q) = Q_{\text{det}}$. $^+$ Means that $\gamma = 1.5$ and $d_{\text{f}} = 5$ was fixed.

Table B2. Maimai: summary of the predictions in the calibration and the validation period made with error models E1–E4 for different temporal resolutions of the hydrological data. Values are medians (and standard deviation) of the quality indices of the deterministic model output for the maximum posterior parameters, as well as those of 500 streamflow realisations produced with the full posterior parameter distributions. Recall that smaller values of Ξ_{reli} and Ω_{spread} indicate better results.

	Calibration									Validation							
Case	Ξ_{reli}	Ω_{spread}	$\widehat{E}_{\text{N,det}}$	E_{N}	$\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	I_{F}	$I_{\text{F,obs}}$	Ξ_{reli}	Ω_{spread}	$\widehat{E}_{\text{N,det}}$	E_{N}	$\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	I_{F}	$I_{\text{F,obs}}$
24 h E1	0.91	0.42	0.92	0.73(0.06)	−8	−8(3.7)	0.77	0.88(0.03)	0.83	0.91	0.4	0.91	0.7(0.08)	−3	−4(3.9)	0.84	0.94(0.04)	0.88
24h E2	0.75	0.52	0.91	0.62(0.1)	−17	−20(7.4)	0.74	0.8(0.03)	0.83	0.93	0.49	0.9	0.59(0.13)	−11	−13(7.9)	0.81	0.85(0.03)	0.88
24 h E3	0.89	0.45	0.91	0.7(0.08)	−11	−11(4.1)	0.79	0.89(0.04)	0.83	0.9	0.43	0.9	0.65(0.1)	−6	−5(4.5)	0.87	0.95(0.04)	0.88
24 h E3a	0.78	0.49	0.91	0.64(0.09)	−16	−18(6.4)	0.75	0.82(0.03)	0.83	0.96	0.47	0.9	0.62(0.11)	−10	−11(6.3)	0.82	0.88(0.04)	0.88
24 h $\widetilde{\text{E4}}$	0.95	0.56	0.92	0.53(0.24)	−6	−16(6.2)	0.81	0.93(0.05)	0.83	0.86	0.54	0.91	0.48(0.29)	−1	−12(6.8)	0.88	0.99(0.05)	0.88
24 h E4	0.98	0.57	0.92	0.6(0.27)	−9	−9(5.5)	0.78	0.9(0.04)	0.83	0.81	0.54	0.91	0.57(0.33)	−4	−3(5.7)	0.85	0.95(0.05)	0.88
24 h $\widetilde{\text{E4a}}$	0.93	0.62	0.92	0.44(0.37)	−5	−24(8.4)	0.8	0.88(0.04)	0.83	0.93	0.63	0.91	0.37(2.74)	−1	−19(10.9)	0.88	0.95(0.05)	0.88
24 h E4a	0.98	0.56	0.92	0.6(0.25)	−10	−10(6.6)	0.77	0.85(0.04)	0.83	0.84	0.56	0.91	0.55(0.94)	−5	−6(7.5)	0.85	0.91(0.05)	0.88
6 h E1	0.91	0.45	0.89	0.69(0.05)	−11	−11(2.2)	0.4	0.63(0.02)	0.46	0.88	0.42	0.88	0.68(0.05)	−5	−5(2.5)	0.45	0.65(0.03)	0.47
6 h E2	0.53	0.67	0.83	0.34(0.19)	−37	−51(9.3)	0.36	0.39(0.01)	0.46	0.73	0.62	0.83	0.37(0.2)	−27	−40(11.2)	0.39	0.43(0.01)	0.47
6 h E3	0.85	0.49	0.84	0.58(0.07)	−17	−18(2.8)	0.45	0.61(0.02)	0.46	0.9	0.46	0.82	0.54(0.09)	−10	−10(3.3)	0.5	0.65(0.03)	0.47
6 h $\widetilde{\text{E3a}}$	0.89	0.56	0.85	0.5(0.12)	−14	−25(6)	0.38	0.45(0.02)	0.46	0.81	0.51	0.84	0.49(0.14)	−6	−16(7)	0.42	0.49(0.02)	0.47
6 h $\widetilde{\text{E4}}$	0.92	0.59	0.89	0.52(0.14)	−13	−14(3.6)	0.42	0.65(0.03)	0.46	0.8	0.56	0.87	0.48(0.21)	−7	−7(4.2)	0.46	0.67(0.03)	0.47
6 h E4	0.97	0.72	0.89	0.39(0.27)	−15	−16(4.1)	0.4	0.65(0.03)	0.46	0.82	0.67	0.87	0.39(0.88)	−9	−9(4.7)	0.44	0.67(0.04)	0.47
6 h $\widetilde{\text{E4a}}$	0.94	0.92	0.89	−0.19(1.76)	−1	−46(12.5)	0.38	0.47(0.02)	0.46	0.86	0.84	0.88	−0.08(1.66)	5	−37(12.2)	0.42	0.51(0.03)	0.47
6 h E4a	0.95	1.45	0.83	−0.45(3.16)	−29	−39(19.3)	0.33	0.44(0.02)	0.46	0.84	1.27	0.83	−0.34(5.2)	−19	−26(20.6)	0.36	0.46(0.03)	0.47
1 h E1	0.91	0.56	0.84	0.48(0.04)	−19	−19(1.2)	0.14	0.56(0.01)	0.13	0.85	0.52	0.8	0.41(0.06)	−11	−12(1.4)	0.15	0.56(0.01)	0.12
1 h E2	0.71	0.87	0.78	−1.19(0.73)	−26	−96(17.2)	0.12	0.13(0)	0.13	0.86	0.77	0.71	−1.38(0.92)	−15	−77(17.1)	0.13	0.14(0)	0.12
1 h E3	0.91	0.58	0.78	0.43(0.04)	−21	−26(1.9)	0.13	0.38(0.01)	0.13	0.71	0.52	0.71	0.33(0.07)	−11	−14(2.3)	0.14	0.39(0.01)	0.12
1 h E3*	0.84	0.5	0.83	0.62(0.02)	−16	−22(2)	0.11	0.32(0.01)	0.13	0.82	0.44	0.8	0.59(0.04)	−6	−10(2.2)	0.12	0.32(0.01)	0.12
1 h E3*†	0.86	0.72	0.78	0.02(0.24)	−30	−51(6.7)	0.11	0.26(0.01)	0.13	0.86	0.65	0.73	−0.14(0.29)	−19	−41(7.3)	0.13	0.24(0.01)	0.12
1 h $\widetilde{\text{E3a}}$*	0.76	0.5	0.81	0.51(0.1)	−20	−36(5.8)	0.11	0.13(0)	0.13	0.86	0.44	0.77	0.43(0.15)	−10	−24(6.2)	0.13	0.15(0)	0.12
1 h $\widetilde{\text{E4}}$	0.88	0.82	0.83	−0.11(0.46)	−16	−34(3)	0.13	0.46(0.02)	0.13	0.7	0.77	0.8	−0.2(1.21)	−9	−25(3.7)	0.14	0.47(0.02)	0.12
1 h $\widetilde{\text{E4}}$*	0.93	0.63	0.86	0.31(0.43)	−8	−24(2.2)	0.12	0.42(0.02)	0.13	0.73	0.59	0.86	0.3(0.56)	−1	−16(2.6)	0.13	0.43(0.02)	0.12
1 h E4*	0.89	1.02	0.8	−0.17(0.61)	−27	−27(3.9)	0.1	0.5(0.02)	0.13	0.67	0.94	0.78	−0.22(0.93)	−16	−16(4.8)	0.11	0.51(0.03)	0.12
1 h $\widetilde{\text{E4a}}$*	0.8	0.4	0.72	0.58(0.33)	18	1(4.3)	0.08	0.12(0)	0.13	0.68	0.4	0.71	0.58(0.27)	19	2(4.8)	0.08	0.12(0.01)	0.12
1 h E4a*	0.92	1.2	0.78	−0.27(1.76)	−35	−42(13.9)	0.1	0.16(0.01)	0.13	0.77	1.08	0.76	−0.32(1.8)	−22	−28(15.7)	0.11	0.17(0.01)	0.12

*: smoothing $P_{\text{err}}(t)$ with a moving-average window of size 5 h before applying Eq. (11). ˜: denotes the option where $\text{mode}(D_Q) = Q_{\text{det}}$. †: $P_{\text{err}} \neq P$.

Appendix C: Specific error models

C1 Normal distribution

$$D_Q = \mathrm{N}(\mu, \sigma)$$
$$\mu(Q_{\mathrm{det}}) = Q_{\mathrm{det}} \quad \sigma(Q_{\mathrm{det}}, a, b, c) = a Q_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + b Q_0,$$
$$\boldsymbol{\psi} = (a, b, c) \tag{C1}$$

Q_0 is a chosen constant to make the fraction that is taken to the power of c non-dimensional. A modification of the constant Q_0 leads to a re-definition of the parameter a. Therefore, introducing the constant Q_0 does not increase the number of parameters but it simplifies the units of the parameters a and b that become the same as those of streamflow, whereas c is non-dimensional. Empirical evidence has shown that the normal distribution works astonishingly well. However, there is still as small number of outliers that violate the distributional assumptions relatively strongly. For this reason, a distribution with heavier tails seems appropriate.

C2 Student's *t* distribution

$$D_Q = T_{d_\mathrm{f},\,\sigma}(\mu, \sigma, d_\mathrm{f})$$
$$\mu(Q_{\mathrm{det}}) = Q_{\mathrm{det}}\, \sigma_{T_{d_\mathrm{f}}} = a Q_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + b Q_0,$$
$$\boldsymbol{\psi} = (a, b, c) \tag{C2}$$

The Student's t distribution with degrees of freedom $d_\mathrm{f} > 2$ is a straightforward candidate with heavier tails that reduces to the normal distribution for $d_\mathrm{f} \to \infty$. Note that we need to rescale the original Student's t distribution, $T(d_\mathrm{f})$, to the standard deviation σ, i.e. $T(\sigma, d_\mathrm{f})$:

$$f_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}(x) = \frac{1}{\sigma}\sqrt{\frac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; f_{\mathrm{T}_{d_\mathrm{f}}}\left(\frac{1}{\sigma}\sqrt{\frac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; x\right) \tag{C3}$$

and

$$F_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}(x) = F_{\mathrm{T}_{d_\mathrm{f}}}\left(\frac{1}{\sigma}\sqrt{\frac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; x\right). \tag{C4}$$

Note that the degrees of freedom, d_f, have to be larger than 2 to make the standard deviation finite and allow for rescaling to a given standard deviation, σ.

C3 Skewed Student's *t* distribution

$$D_Q = \mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}](Q_{\mathrm{det}}, \sigma, d_\mathrm{f}, \gamma)$$
$$\sigma_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]} = a Q_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + b Q_0,\ \boldsymbol{\psi} = (a, b, c) \tag{C5}$$

To account for the often encountered case of skewed errors of deterministic hydrological models, we transform the Student's t distribution with a generally applicable method for skewing distributions (Fernandez and Steel, 1998). For $\gamma = 1$, the skewed Student's t distribution reduces to the conventional Student's t distribution. Note that the skewing happens after we rescaled the original Student's t distribution to the standard deviation σ. The skewing changes the distributions' standard deviation again, thus $\sigma \neq \sigma_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]}$. The density and cumulative distribution functions of the skewed rescaled distribution, are as follows:

$$f_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]}(x) = \begin{cases} \dfrac{2}{\gamma + \frac{1}{\gamma}} f_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}(\gamma x) = \dfrac{2}{\gamma + \frac{1}{\gamma}} \dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}} \\ \quad f_{\mathrm{T}_{d_\mathrm{f}}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; \gamma x\right) \quad \text{if } x \le 0, \\ \\ \dfrac{2}{\gamma + \frac{1}{\gamma}} f_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}\left(\dfrac{x}{\gamma}\right) = \dfrac{2}{\gamma + \frac{1}{\gamma}} \dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}} \\ \quad f_{\mathrm{T}_{d_\mathrm{f}}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; \dfrac{x}{\gamma}\right) \quad \text{if } x \ge 0. \end{cases} \tag{C6}$$

and

$$F_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]}(x) = \begin{cases} \dfrac{2}{1+\gamma^2} F_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}(\gamma x) \\ \quad = \dfrac{2}{1+\gamma^2} F_{\mathrm{T}_{d_\mathrm{f}}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; \gamma x\right) \text{ if } x \le 0, \\ \\ \dfrac{1}{1+\gamma^2} + \dfrac{2}{1+\frac{1}{\gamma^2}}\left(F_{\mathrm{T}_{d_\mathrm{f},\,\sigma}}\left(\dfrac{x}{\gamma}\right) - \dfrac{1}{2}\right) \\ \quad = \dfrac{1}{1+\gamma^2} + \dfrac{2}{1+\frac{1}{\gamma^2}}\left(F_{\mathrm{T}_{d_\mathrm{f}}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{d_\mathrm{f}}{d_\mathrm{f}-2}}\; \dfrac{x}{\gamma}\right) - \dfrac{1}{2}\right) \\ \quad \text{if } x \ge 0. \end{cases} \tag{C7}$$

And the mean and the variance of the skewed rescaled distribution are as follows:

$$\mu_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]} = 2\sigma\, \frac{\gamma^2 - \frac{1}{\gamma^2}}{\gamma + \frac{1}{\gamma}}\, \frac{\sqrt{d_\mathrm{f}(d_\mathrm{f}-2)}}{d_\mathrm{f}-1}\, \frac{\Gamma\left(\frac{d_\mathrm{f}+1}{2}\right)}{\sqrt{\pi\, d_\mathrm{f}}\, \Gamma\left(\frac{d_\mathrm{f}}{2}\right)} \tag{C8}$$

and:

$$\sigma^2_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]} = \frac{\gamma^3 + \frac{1}{\gamma^3}}{\gamma + \frac{1}{\gamma}}\, \sigma^2 - \mu^2_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\sigma}]}$$
$$= \left(\frac{\gamma^3 + \frac{1}{\gamma^3}}{\gamma + \frac{1}{\gamma}} - 4\left(\frac{\gamma^2 - \frac{1}{\gamma^2}}{\gamma + \frac{1}{\gamma}}\right)^2 \frac{d_\mathrm{f}(d_\mathrm{f}-2)}{(d_\mathrm{f}-1)^2}\, \frac{\Gamma^2\left(\frac{d_\mathrm{f}+1}{2}\right)}{\pi\, d_\mathrm{f}\, \Gamma^2\left(\frac{d_\mathrm{f}}{2}\right)}\right)\sigma^2. \tag{C9}$$

To shift the distribution we can evaluate

$$f_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\,\sigma}]}(x - Q_{\mathrm{det}}), \tag{C10a}$$

$$f_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\sigma}]}(x + \mathrm{med}_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\sigma}]} - Q_\mathrm{det}), \tag{C10b}$$

$$f_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\sigma}]}(x + \mu_{\mathrm{sk}_\gamma[\mathrm{T}_{d_\mathrm{f},\sigma}]} - Q_\mathrm{det}). \tag{C10c}$$

In these cases, the mode, the median and the mean are located at x_0, respectively.

Appendix D: Notation

P	Precipitation used as an input to the hydrological model.
P_{err}	Precipitation used as an input to the error model where needed (not to the hydrological model).
$Q_{\mathrm{det}}(t, \boldsymbol{\theta})$	Deterministic hydrological model providing streamflow as a function of time, t, and hydrological model parameters $\boldsymbol{\theta}$.
$\widehat{Q}_{\mathrm{det}}$	Deterministic hydrological model output corresponding to the parameter vector $\widehat{\boldsymbol{\theta}}$ with the maximum posterior density.
$Q_{\mathrm{obs}}(t)$	Observed streamflow at time t.
$Q_{\mathrm{trans}}(\eta)$	Function transforming η into streamflow (used to sample from the probabilistic model consisting of the hydrological model and the error model).
D_Q	Distribution of observed streamflow at a certain point in time, given the output of the deterministic hydrological model at the same point in time.
$\boldsymbol{\theta}$	Parameters of the deterministic hydrological model, Q_{det}.
$\boldsymbol{\psi}$	Parameters of the error model, including heteroscedasticity and correlation parameters.
η	Autocorrelated, stochastic process with standard normal asymptotic distribution that serves to describe the autocorrelation of the errors of the deterministic hydrological model.
τ	Characteristic correlation time of the process η.
τ_{min}	Minimum value of τ in the cases where τ is a function of P_{err} and therefore of time.
τ_{max}	Maximum value of τ in the cases where τ is a function of P_{err} and therefore of time.
F_X	Cumulative distribution function of the distribution X.
f_X	Probability density function of the distribution X.
$E[X]$	Expected value of the random variable X.
$\mathrm{N}(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ.
$\mathrm{T}(d_{\mathrm{f}}, \sigma)$	Rescaled Student's t distribution with d_{f} degrees of freedom and standard deviation σ.
$\mathrm{SKT}(\mu, \sigma, d_{\mathrm{f}})$	Shifted and rescaled skewed Student's t distribution with mean μ, standard deviation σ and d_{f} degrees of freedom.
I_{F}	The median of the flashiness indices of all the individual model realisations constituting a sample of model outputs.
$\widehat{I}_{\mathrm{F,\,det}}$	The flashiness index of $\widehat{Q}_{\mathrm{det}}$.
$I_{\mathrm{F,\,obs}}$	The flashiness index of Q_{obs}.
E_{N}	The median of the Nash–Sutcliffe efficiencies (Nash and Sutcliffe, 1970) of all the individual model realisations constituting a sample of model outputs.
$\widehat{E}_{\mathrm{N,\,det}}$	The Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) of $\widehat{Q}_{\mathrm{det}}$.
Δ_Q	The median of the relative errors in cumulative streamflow of all the individual model realisations constituting a sample of model outputs.
$\widehat{\Delta}_{Q,\,\mathrm{det}}$	The relative error in cumulative streamflow of $\widehat{Q}_{\mathrm{det}}$.
Ξ_{reli}	Reliability metric; the complement of the reliability metric defined in McInerney et al. (2017).
Ω_{spread}	Relative spread metric; equal to the precision metric defined in McInerney et al. (2017).
OU process	Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/hess-23-2147-2019-supplement.

Author contributions. PR conceptualized the general theory with contributions from LA and FF. LA developed the conceptual adaptations and improvements of the suggested approaches with contributions from PR and FF. All authors designed the experiments and FF and LA selected the test cases. LA did the implementation, data compilation, and testing. LA wrote the paper with contributions from FF and PR.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study was funded by the Swiss National Science Foundation (grant 200021_163322). The authors thank MeteoSwiss (Swiss Federal Office of Meteorology and Climatology) for the meteorological data concerning the Murg catchment, Massimiliano Zappa for the preprocessing of this data and Jeffrey McDonnell for the hydrological data of the Maimai catchment. Lorenz Ammann thanks Omar Wani for the inspiring discussions and exchange of ideas. Dmitri Kavetski provided valuable feedback on a draft of this paper. The authors also thank Alberto Montanari, Jasper Vrugt and two anonymous referees for their feedback and their help in improving this paper.

Review statement. This paper was edited by Erwin Zehe and reviewed by Jasper Vrugt and two anonymous referees.

References

Baker, D. B., Richards, R. P., Loftus, T. T., and Kramer, J. W.: A new flashiness index: characteristics and applications to midwestern rivers and streams, J. Am. Water Resour. As., 40, 503–522, https://doi.org/10.1111/j.1752-1688.2004.tb01046.x, 2004.

Bárdossy, A. and Das, T.: Influence of rainfall observation network on model calibration and application, Hydrol. Earth Syst. Sci., 12, 77–89, https://doi.org/10.5194/hess-12-77-2008, 2008.

Bates, B. C. and Campbell, E. P.: A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, Water Resour. Res., 37, 937–947, https://doi.org/10.1029/2000wr900363, 2001.

Bertuzzo, E., Thomet, M., Botter, G., and Rinaldo, A.: Catchment-scale herbicides transport: Theory and application, Adv. Water Resour., 52, 232–242, https://doi.org/10.1016/j.advwatres.2012.11.007, 2013.

Beven, K. and Westerberg, I.: On red herrings and real herrings: disinformation and information in hydrological inference, Hydrol. Process., 25, 1676–1680, https://doi.org/10.1002/hyp.7963, 2011.

Boltz, S., Debreuve, E., and Barlaud, M.: kNN-based high-dimensional Kullback-Leibler distance for tracking, in: Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2007), IEEE, 6–8 June 2007, Santorini, Greece, https://doi.org/10.1109/wiamis.2007.53, 2007.

Brammer, D. D. and McDonnell, J. J.: An Evolving Perceptual Model of Hillslope Flow at the Maimai Catchment, Advances in hillslope processes, 1, 35–60, 1996.

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, J. Hydrol., 298, 242–266, https://doi.org/10.1016/j.jhydrol.2004.03.042, 2004.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrol. Earth Syst. Sci., 17, 4209–4225, https://doi.org/10.5194/hess-17-4209-2013, 2013.

Duan, Q., Sorooshian, S., and Ibbitt, R. P.: A maximum likelihood criterion for use with data collected at unequal time intervals, Water Resour. Res., 24, 1163–1173, https://doi.org/10.1029/wr024i007p01163, 1988.

Evin, G., Kavetski, D., Thyer, M., and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour. Res., 49, 4518–4524, 2013.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resour. Res., 50, 2350–2375, https://doi.org/10.1002/2013wr014185, 2014.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, 1–13, https://doi.org/10.1029/2010wr010174, 2011.

Fenicia, F., Kavetski, D., Reichert, P., and Albert, C.: Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Empirical Analysis of Fundamental Properties, Water Resour. Res., 54, 3958–3987, https://doi.org/10.1002/2017wr021616, 2018.

Fernandez, C. and Steel, M. F. J.: On Bayesian Modeling of Fat Tails and Skewness, J. Am. Stat. Assoc., 93, 359–371, 1998.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J.: emcee: The MCMC hammer, Publ. Astron. Soc. Pac., 125, 306–312, 2013.

Freer, J., Beven, K., and Ambroise, B.: Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach, Water Resour. Res., 32, 2161–2173, https://doi.org/10.1029/95wr03723, 1996.

Hannachi, A.: Intermittency, autoregression and censoring: a first-order AR model for daily precipitation, Meteorol. Appl., 21, 384–397, https://doi.org/10.1002/met.1353, 2012.

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, 1–19, https://doi.org/10.1029/2011wr010748, 2011.

Kavetski, D., Franks, S. W., and Kuczera, G.: Confronting input uncertainty in environmental modelling, in: Water Science and Application, 49–68, American Geophysical Union, San Francisco, USA, https://doi.org/10.1029/ws006p0049, 2003.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water

Resour. Res., 42, 1–9, https://doi.org/10.1029/2005wr004368, 2006.

Kloeden, P. E. and Platen, E.: Numerical Solution of Stochastic Differential Equations, Springer, Berlin, 1995.

Krzysztofowicz, R.: Bayesian system for probabilistic river stage forecasting, J. Hydrol., 268, 16–40, https://doi.org/10.1016/s0022-1694(02)00106-3, 2002.

Kuczera, G.: Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, Water Resour. Res., 19, 1151–1162, https://doi.org/10.1029/wr019i005p01151, 1983.

Kuczera, G. and Franks, S.: Testing hydrologic models: Fortification or falsification?, in: Mathematical Models of Large Watershed Hydrology, edited by: Singh, V. P. and Frevert, D. K., Water Resources Publications, Highlands Ranch, Colorado 80163-0026, USA, 2002.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, Ann. Math. Stat., 22, 79–86, https://doi.org/10.1214/aoms/1177729694, 1951.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998wr900018, 1999.

McGlynn, B. L., McDonnell, J. J., and Brammer, D. D.: A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand, J. Hydrol., 257, 1–26, 2002.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resour. Res., 53, 2199–2239, https://doi.org/10.1002/2016wr019168, 2017.

MeteoSwiss: https://www.meteoschweiz.admin.ch/home/service-und-publikationen/beratung-und-service/datenportal-fuer-experten.html (last access: 18 April 2019), 2018.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, Science, 319, 573–574, https://doi.org/10.1126/science.1151915, 2008.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, 1–11, https://doi.org/10.1029/2003wr002540, 2004.

Montanari, A. and Koutsoyiannis, D.: Modeling and mitigating natural hazards: Stationarity is immortal!, Water Resour. Res., 50, 9748–9756, https://doi.org/10.1002/2014wr016092, 2014.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, https://doi.org/10.1029/2009WR008328, 2010.

Schaefli, B., Talamba, D. B., and Musy, A.: Quantifying hydrological modeling errors through a mixture of normal distributions, J. Hydrol., 332, 303–315, https://doi.org/10.1016/j.jhydrol.2006.07.005, 2007.

Scharnagl, B., Iden, S. C., Durner, W., Vereecken, H., and Herbst, M.: 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, https://doi.org/10.5194/hessd-12-2155-2015, 2015.

Schleppi, P., Waldner, P. A., and Fritschi, B.: Accuracy and precision of different sampling strategies and flux integration methods for runoff water: comparisons based on measurements of the electrical conductivity, Hydrol. Process., 20, 395–410, https://doi.org/10.1002/hyp.6057, 2006.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, W10531, https://doi.org/10.1029/2009WR008933, 2010.

Seibert, J. and McDonnell, J. J.: On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, Water Resour. Res., 38, 1241, https://doi.org/10.1029/2001WR000978, 2002.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., and Sisson, S.: Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, Water Resour. Res., 46, 1–11, https://doi.org/10.1029/2010wr009514, 2010.

Sun, X., Mein, R., Keenan, T., and Elliott, J.: Flood estimation using radar and raingauge data, J. Hydrol., 239, 4–18, https://doi.org/10.1016/s0022-1694(00)00350-4, 2000.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, 1–22, https://doi.org/10.1029/2008wr006825, 2009.

Uhlenbeck, G. E. and Ornstein, L. S.: On the Theory of the Brownian Motion, Phys. Rev., 36, 823–841, https://doi.org/10.1103/physrev.36.823, 1930.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, Environ. Modell. Softw., 24, 1209–1222, https://doi.org/10.1016/j.envsoft.2009.04.001, 2009.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, https://doi.org/10.5194/hess-5-13-2001, 2001.

Wani, O., Scheidegger, A., Cecinati, F., Espadas, G., and Rieckermann, J.: Exploring an alternative to additive error models – for non-negative and autocorrelated time series in hydrology, in preparation, 2019.

Yang, J., Reichert, P., Abbaspour, K. C., and Yang, H.: Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, J. Hydrol., 340, 167–182, 2007.

Zeger, S. L. and Brookmeyer, R.: Regression Analsis with Censored Autocorrelated Data, J. Am. Stat. Assoc., 81, 722–729, https://doi.org/10.2307/2289003, 1986.