Author's response

Note that all the references to the manuscript (made by referees and by authors) refer to the original version of the manuscript, not the revised markup version contained in this file.

RC: Referee comment

AC: Author comment

Changes made to the manuscript are highlighted in italic font.

Reply to Referee #1, Jasper Vrugt

J. Vrugt: Summary: In this paper the authors introduce a parametric framework to residual analysis. This approach leads to formulation of a likelihood function which, with a suitable prior distribution, helps to evaluate the posterior density of nontraditional residual time series, e.g. truncated and subject to various degrees of skew, kurtosis and serial correlation. The framework allows for the use of transient nuisance variables (hyper parameters) to help accommodate so-called non-stationary residual patterns. The framework presented herein differs a bit from the standard likelihood paradigm in that the starting point is some parametric family of distributions which describes the likelihood of observing the data, Q, given current model output, Qdet. Authors claim that the proposed likelihood function improves probabilistic inference of hydrologic models via MCMC machinery – with a more realistic description of parameter and predictive uncertainty. I enjoyed reading this paper as it combines theory development with practical application. The paper is well written and should be of interest to the readership of HESS. I hope the authors consider the following comments – I believe those will help to further improve the quality of this manuscript. Note, comments appear in order of my reading of the paper.

> **AC:** Thanks for this general feedback. We have to clarify that the ability to deal with non-stationary correlation (or other parameters) is independent of the presented likelihood framework (as referee #2 correctly pointed out), and could (and should) also be achieved with other frameworks / methods.
>
> *We included a corresponding sentence in Sect. 5.2.*.

J. Vrugt (1): Page 5, Line 9-11. Authors state that most (many) modelers will have an intuitive idea about the probability distributions of the observations for a given model output. I disagree with this assertion. For the sake of my argument, lets follow the hydrologic example as presented in this work. Let's assume that the model simulates a discharge of 20 mm/day. What would be a reasonable expectation of the actual (observed) discharge at that time? 15? 30? I cannot confidently claim that I would know what probability distribution to assume for the observed discharge at that time. Of course, if 20 mm/day is among the largest simulated values, then I would generally expect the dispersion of this supposed distribution to be larger than for a simulated value of 5 mm/day. Yet, this is only the dispersion – I would not really have an idea about the underlying distribution – would I center this distribution on 20 mm/day? Or is my model systematically under or overestimating the data so that I should shift the distribution to higher or lower values, respectively. Of course, for low discharge values I know that the

distribution is truncated at zero – and probably has a tail to the right. But then again do I center the distribution on the model simulated value? Or do we shift it up or downward? In other words, I do not agree with the assertion that many modelers will have an intuitive idea what the distribution of the observed discharge would be if the model output is known.

> **AC (1)**: This is an interesting point of discussion about one of the main motivations for the presented likelihood framework. Interestingly, there is a contrasting opinion of the referee J. Vrugt and the author of a short comment, Alberto Montanari, on exactly this point. We acknowledge that our wording "many modellers will have an intuitive idea about the probability distribution ..." is too strong. We agree that it can be difficult to formulate this distribution of the observed streamflow, as the example of J. Vrugt shows.
>
> *We rephrased the sentence accordingly.*
>
> However, in case we do have at least some idea about the shape of the distribution, the presented framework allows to incorporate this as prior knowledge. If we have no idea about the distribution, the presented framework is still useful, because we can communicate and discuss our assumptions in the space of streamflow (with the corresponding units), as the example in the comment of the referee shows. With previously used approaches to deal with skewness and kurtosis (e.g. Box-Cox transformation, generalized likelihood function) it is more difficult to discuss these assumptions, because they are made in transformed (Box-Cox) or innovation (generalized likelihood) spaces, which are less intuitive for us. Our point was that it is easier for hydrologists (although admittedly still not easy) to discuss the marginal distribution of streamflow (because they have been confronted with deviations of model results from observations for this quantity in the past) rather than Box- Cox parameters or the distribution of innovations without simple access to the consequences on marginal streamflow distributions. In summary, this discussion illustrates the major advantage of the presented framework: that distributional assumptions are transparent and easy to communicate, which means that they can be better discussed and questioned.
>
> *We mentioned this shortcoming of previous approaches more explicitly in the paragraph on page 3, line 20. Accordingly, we expanded the first paragraph of Sect. 2.1 to provide a clearer motivation for the presented approach of parameterizing the distribution of streamflow given model output, as compared to transformation approaches (Box-Cox) or probabilistic models formulated in the innovation space (generalized likelihood).*

J. Vrugt (2): Page 5, Line 22-23. The authors refer to Eq. (3) before presenting Eq. (2). Do not understand why this is done – would think that text can be presented so that Eq. (3) follows first – then followed by Eq. (2). Note, is Eq. (3) needed after all? The right-hand-side of Eq. (3) can be placed at end of Eq. (2) – then the index needs to be fixed.

> **AC(2):** We agree to reverse the order of the equations. Equation (3) is very important to introduce the transformation function before it is applied to the actual time series, as this transformation is the key of our concept of introducing autocorrelation for arbitrary marginal discharge distributions.

We reversed the order of Eq.(2) and Eq.(3) and we edited the adjacent paragraphs to provide better explanations of the equations and the idea of the transformation to combine autocorrelation with arbitrary marginal distributions of streamflow.

J. Vrugt (3): Page 5, Line 27-29: I do not understand the statement that truncation at zero would lead to lighter tails on the lower end. Yes, truncation would move the probability of negative streamflow values to streamflow values larger than zero. In essence, one could then argue that the tail at the right-hand-side may become larger – as the pdf has to integrate to unity. Yet, because of truncation the left tail is essentially gone if simulated streamflow values are close to zero. The wording "lighter tails" may be a bit confusing as the tail is truncated. It is no longer there.

AC(3): It is true that the negative part of the distribution DQ is truncated at each individual time step, so the negative tail at each time step is no longer there. However, here we refer to the marginal distribution of eta over all time steps, and usually there will be no sharp "cut" visible, since the truncation happens at different values at each time step.

We included a corresponding sentence. *We also more clearly discussed that our framework allows for truncation with compensation by increased density for positive values (as described by the referee) or for assigning a finite probability for an observed discharge of zero (as actually done in this study).*

J. Vrugt(4): Page 5, Eq. (2) – (3) – thus, eta is the normally transformed counterpart of Q – with truncation accounted for?

AC(4): Yes, this is exactly right. Together with the changes made regarding comment 3, we hope that this became clearer. Truncation will only be needed if the distributional shape of the discharge extends to negative values. This may not always have to be the case.

J. Vrugt(5): Equation (4) – authors may consider for normal distribution, N, instead \mathcal(N)(a,b), where "a" (mean) is the first term between brackets in Eq. (5) and "b" is the second term in Eq. (4). In text below Eq. (4) authors could then explain that "a" is the mean of the distribution and b is the variance.

AC(5): We agree that it must be made more explicit that the first term is the mean and the second is the standard deviation.

Rather than introducing two new variables, we stated in the text that the two elements are the mean and the standard deviation.

J. Vrugt(6): Eq. (6) – reference should be given.

AC(6): To clarify the derivation, we replaced the paragraph around Eqs. (5) and (6) by:

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

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

with

$$\phi = \exp\left(-\frac{\Delta t}{\tau}\right) \quad \text{or} \quad \tau = -\frac{\Delta t}{\log(\phi)}$$

This is an AR1 process with autoregression coefficient ϕ and white noise variance $(1-\phi^2)$.

J. Vrugt(7): Page 6, Line 12-14. Maybe I am missing something here, but with any other likelihood function one can ignore missing data as well? One simply does not include this particular observation in the likelihood function. The authors may have a point if serial correlation is considered – then this removal is not straightforward as it breaks the AR-error model.

> **AC(7):** Yes, we agree. Any likelihood can deal with missing data when neglecting correlation, but it requires more effort with an AR error model. Since we think that considering correlation is important, we think it is necessary that future likelihoods can accommodate both, correlation and missing data (or varying time step sizes) naturally. Our point is that this is particularly simple in the suggested approach as it does not need any changes because there is no underlying assumption of equidistant points in time.
>
> *We mentioned this more explicitly in the manuscript*.

J. Vrugt(8): Eq. (7) – top line of curly brace may fit on one line if authors define rho = (ti+1 – ti)/tau, and then use rho in the equation – maybe etatrans written as etaT.

> **AC(8):** We agree that Eq. (7) is not ideally displayed. We prefer to implement the latter proposition of the referee.
>
> *We replaced eta_trans by eta(ti), i.e. we substituted Eq. (2) into Eq. (7). Since the dependence of eta(t_i) on Q(t_i) is then not explicit anymore, we added a statement about that dependence and referred to Eq. (2). Making these changes, we realized that "theta" is not properly introduced in this section and the dependence of Qdet on "theta" is not consistently stated. We introduced "theta" and added the dependence at two points in the text. We also realized that Eq. (2) still contained a "xi"-function from a previous notation version and removed it from Eq. (2) and from the Appendix.*

J. Vrugt(9): Then notation – not sure about the guidelines of HESS, but should theta (parameter vector) not be upright-bold instead of italic-bold? Same holds for the nuisance variables, psi.

> **AC(9):** The current guidelines of HESS are italic bold for vectors, according to the information we have.

J. Vrugt(10): Is notation DQ required or would fQ suffice instead? Then, the text would talk about a distribution of Q – instead of DQ.

> **AC(10):** This would be a possibility, and it would probably make the equations better readable. However, talking about the "distribution of Q" instead of DQ, would make the text quite a bit

longer, since the term appears often. We would prefer to stay with the name DQ, because with think it is overall simpler to read.

J. Vrugt(11): A limitation of Eq. (4) is that serial correlation at higher-order lags cannot be modelled, right? Unless you specify different "rho's" in Eq. (6) – but this then leads to multiple likelihoods. This limitation should be stated in the text as residuals may exhibit/show residual correlation beyond lag-1.

AC(11): Yes, we fully agree with this comment.

We included a corresponding statement about Eq. (4) in this version of the manuscript.

J. Vrugt(12): In Eq. (8) how do we compute the first term on the right-hand-side – that is – the likelihood of the zeroth discharge observation (at t0)? Do we assume normality with dispersion of variance/(1-rhoˆ2)?

AC(12): This term is calculated with Eq.(1). We recognize that it is confusing that the index "i" refers to the current time step for which we want to calculate the likelihood in Eq.(8), but that it refers to the time step before the current time in Eq. (7).

We referred explicitly to Eq. (1) and also modified the index "i" in Eq. (7), in the adjacent text, as well as in the Appendix A, so that it has the same meaning as in Eq. (8). We also changed the time index in section 2.3, so that it is consistent with the rest of the manuscript.

J. Vrugt(13): Page 7, Line 12-13: The statement "the likelihood function can be evaluated analytically" is a bit confusing to me. What does the word "analytical" mean in this context? Most other commonly used likelihood functions in the applied (hydrologic) literature are simple to evaluate in practice, right? That means numerically. All that is needed are the model output and the data? What is different in the present context?

AC(13): We agree that this is a property shared by most likelihoods formulated on top of a deterministic hydrological model. We wanted to express that our framework still belongs to that class and does not lead to additional numerical effort as e.g. stochastic hydrological models that may require PMCMC or ABC rather than standard MCMC. It was not our intention to state that our model is special in this respect.

We clarified this and replaced the expression "evaluated analytically" with "available in closed form" to make it clearer what we mean here.

J. Vrugt(14): The authors use the affine invariant ensemble sampler of Foreman and Mackay et al. (2013) to sample the posterior parameter and nuisance variable distribution. The article would benefit from some more background information – that is – algorithmic settings (number of walkers, the types of moves that are considered, etc.). Note, that this ensemble sampler has many elements in common with the DREAM family of MCMC algorithms – which uses parallel direction and snooker moves. For later work it may be interesting to compare both methods in terms of efficiency – and to evaluate the power and usefulness of the walk, stretch and replacement move. Note, that the ensemble sampler has two important shortcomings; 1) detailed balance requires the use of a relatively large number of walkers

(chains) – this is a significant disadvantage for higher dimensional problems as each chain needs burn-in before reaching the target distribution, and 2) the walkers require stepwise updating – this guarantees reversibility but does not make the sampler amenable to distributed computing, wherein each chain is evolved on a different core/node.

AC(14): We agree that more background information should be provided on this.

We included the specific settings used for sampling with this ensemble sampler.

We also agree that it would be interesting to compare the performance of the sampler applied in this study and the DREAM samplers in a future study.

J. Vrugt(15): Equation (10) – the subscript "F" in the flashiness index, should this not be regular font – that is – upright? As "F" is an abbreviation for "flashiness" and not a variable. Same holds for some of the other summary metrics used in this paper, for example the Nash-Sutcliffe efficiency (subscript "N" should be regular = upright font). Note, that on Page, 8, Line 25 correct notation is used for the flashiness index of the deterministic model output.

AC(15): This is right, thanks for the notice.

We checked and improved regular versus italics fonts in equations throughout the manuscript. Based on this we found that the vector notation (bold) for streamflow was not consistent in Sect. 2.3. Therefore we introduced the vector notation for time series of streamflow at the beginning of the section and improved the notation of the equations in that section.

J. Vrugt(16): Page 5, Line 24: "maximum posterior parameter values" – this is rather awkward wording as it literally means – the largest posterior parameter values. And it is not clear what this means either as each dimension of the target distribution will have a maximum posterior value – but all these maxima combined are unlikely to make up an actual posterior sample. Instead, what the authors should use is "maximum a-posteriori density (MAP) parameter values" – that is – the parameter values that maximize the product of the prior density and the likelihood.

AC(16): We assume that the referee means Page 8, Line 24 instead of Page 5, Line 24. What we mean by this is the single parameter vector that is associated with the largest posterior probability density of all the points in the parameter sample. As we are not referring to marginal posterior densities, this can hardly be misunderstood in the way the referee argues. However it certainly makes sense to add the word "density" to "maximum posterior".

We changed the wording "maximum posterior parameter values" to "parameter values at the maximum posterior density"

J. Vrugt(17): Eq. (15) and (16) list the flux and water balance equations used by the hydrologic model – but equally important what numerical solution method is used to solve these equations? I assume that the authors have used an implicit solution with time-variable integration step? Solution maintains mass balance?

AC(17): We very much agree with the referee. This information should be provided.

We included a reference to the software used to implement the hydrological model, which also contains information about the numerical integration schemes.

J. Vrugt(18): Page 12, Line 5: Why are these model parameters held constant? Why are they not part of the inference – this would be much stronger in my view. If held constant, then how does one know the assumed values are reasonable for the catchment of interest? Note, if I look at the equations then m, alpha and beta must have a large impact on the simulated model output. Hence, unless these parameters have a strong physical underpinning I do not see why one would keep them fixed in the present work. Certainly, the values of m, alpha and beta will affect the residual analysis.

AC(18): We agree that in principle, it is always desirable to infer more parameters. The mentioned parameters were kept fixed to keep the hydrological model parsimonious. Fixing some of the parameters is commonly done in hydrological bucket models, for example, the widely used GR4J model has 4 parameters that are inferred, which is equal to the number of hydrological parameters we infer in this study, and it has other parameters that are kept fixed, including the parameter that is equivalent to "beta" in this study. "m" can be seen as a smoothing parameter, and m=0.01 means that there is close to full evaporation as long as the reservoir Su is not empty. "alpha=2" was found to lead to reasonable results in both the investigated catchments and was fixed because of its potential interactions with kf. We do admit that we do not know if the fixed values of "beta" and "m" are ideal for the investigated catchments. Since we reached good fits with at least some error models in both catchments, we would argue that the values of "beta" and "m" are proven to be reasonable. Often when applying a hydrological model to a catchment, we do not really know whether the model is perfectly appropriate for that catchment and we cannot infer all the (potentially many) parameters of the model. Also, systematic errors are common in practice, so we do not want to avoid them here by overly complex models. One could argue that this limits the transferability of the results to other, more complex models. One could also argue that we should have tested different hydrological models, more catchments and more temporal resolutions to obtain more generalizable results. However, the focus of this paper is on the method development, which allows only for a limited amount of application case studies and comparisons.

We included the above mentioned explanations as to why those parameters were kept fixed, but we did not additionally include model runs where those parameters are fitted.

J. Vrugt(19): The authors do not consider highly relevant work by Scharnagl et al. (2015) published in HESS: Inverse modeling of in situ soil water dynamics: accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals. This work also used a Student distribution for the conditional density of the residuals – and combined this with the template function of Fernandez and Steel (1998) to enable treatment of skewed residual distributions. Given the similarities with the work presented in this paper I think it is important for the authors to consider the listed work of Scharnagl et al.

AC(19): We agree that the work of Scharnagl et al. is related to the topic of this study and we were not aware of it, since it was not published as a final paper in HESS. Their "Likelihood 2" uses a skewed Student t-Distribution, but they use it to describe the probability density of the

innovations, like Schoups and Vrugt (2010), not the probability density of the observed streamflow, as is done in this study. A difference to Schoups and Vrugt (2010) is that Scharnagl et al. (2015) apply the autocorrelated process to the standardized residuals, as the correction suggested by Evin et al. (2013). However, this approach does not give satisfying results in that case. Then, the relevance of "Likelihood 3" in Scharnagl et al. (2015) for predictive application was correctly questioned by one of the referees.

We included a reference to that discussion in the manuscript in the introduction and in Sect. 2.1.

J. Vrugt(20): Eq. (18) – does this function satisfy the laws of total expectation and total variance? This is a concern not typically addressed in the hydrological literature – but the paper by Hernandez-Lopez in HESS (2017) makes some important points regarding preservation of expectation and variance of the error model.

AC(20): The Law of Total Expectation and the Law of Total Variance are statistical theorems. There is no way of violating them for any correctly formulated probabilistic model. We are formulating a joint probability density of discharge at all observations points in equation (8) conditional on the output of the deterministic model. The choice and parameterization of the discharge distribution does not change the validity of fundamental statistical theorems. For this reason, the consideration of heteroscedasticity by Eq. (18) cannot lead to a violation of the Total Laws. Note that we carefully transform the distribution assumed for "eta" to the distribution of "Q" in equation (7); not doing this carefully could be a potential source of error and could lead to a violation of statistical theorems.

Why do Hernandez-Lopez (2015) state that the fulfillment of statistical theorems must be guaranteed by eliminating parameters from MCMC sampling and calculating them from the other components of the sample point (section 4.4 in their paper)? This argument is based on a fundamental misinterpretation of a statistical equation that is valid, if correctly interpreted. Their derivation of equation (22) resp. (B9) in appendix B demonstrates, that this equation links the parameters α and κ, the error variance, the discharge variance and expectation for an error model with fixed parameters α and κ (see equation B5 where this assumption is used). In Bayesian inference, α and κ become random variables and equation (22) is no longer valid (it would contain a sum of random and non-random variables [the expectation and the variance of a random variable are not random]). Applying this invalid equation is the first problem of their approach. The second problem is that the Laws of total Expectation and Variance are integral equations over a multivariate distribution. They have no meaning for individual sampling points to which they apply them. The full sample will fulfill the statistical theorems as a result of the consistency of the approach and without explicit enforcement.

The more interesting question is whether the expectation of the probabilistic model for a given deterministic model output is equal to this deterministic model output. Our framework makes the formulation of such models possible (e.g. a lognormal distribution with mean equal to the deterministic model output). This seems at the first sight a desirable property of the model as it guarantees mass conservation (if the deterministic model conserves mass). Unfortunately, our

experience with such error model formulations were unsatisfactory. In cases in which the model output is very small, even small observations errors can lead to observations that are orders of magnitude larger than the output of the deterministic model and would thus require an extremely strongly skewed distribution. The consequence of such extremely skewed distributions would be that for each "large observation" a very large number of very small observations would be needed to keep the mean (as these observations cannot be much smaller than a small output of the hydrological model). In our experience, such distributions lead to unsatisfactory fits. Thus, the non-negativity of discharge observations (for non-tidal rivers) makes it practically nearly impossible to keep mass balances at very low discharge if there is a considerable observation error.

We added a paragraph in Sect. 5.4 to mention this problem which may also not have gained sufficient attention in the literature.

As for the Law of Total Expectation and Variance, we felt it unnecessary to state the fulfillment of any laws of probability in the paper as this is a property of any correctly formulated model.

J. Vrugt(21): I am wondering whether readability of the paper would improve if the section on error models is placed directly after the likelihood section. Indeed, the likelihood contains tau – which is then defined (among others) in the error model section.

AC(21): We agree with this suggestion.

We changed the order of the sections in the manuscript. In addition, we changed the first sentence of the section on error models slightly, so that it fits better to the new position in the manuscript. The paragraph on the priors, which used to be in the section about error models, was transformed into a separate section (3.3), and some wordings were changed in that paragraph.

J. Vrugt(22): Page 11, Line 16: What has happened to the index time in the formulation of Qdet? It appears on the left-hand side but does not appear on the right-hand side. Also, what are Qs and Qf? These entities are introduced but they are not discussed nor do they appear elsewhere in the paper?

AC(22): We agree that the arguments "t" and "θ" should also appear on the right hand side of the equation and that Qs and Qf should be mentioned in the text. They are the fast and the slow flow components of the model, respectively, and are given by Eq. (15) and illustrated in Fig. 1.

We added the arguments "t" and "tau" on the right hand side and included a statement about Qs and Qf in the adjacent text.

J. Vrugt(23): At this point I am wondering why the authors are not using the more common terminology of P(.) for prior distribution and L(.|.) for likelihood function.

AC(23): Only when the output of the probabilistic model is replaced by the observed data for inference, we obtain the likelihood as a function of the parameters given the observed data. The likelihood function is therefore crucial for inference. It is hardly possible to formulate this function directly. This is why scientists formulate probabilistic models as probability distributions

of outcomes given parameters and only afterwards get the likelihood function by substituting the observations for the outcomes. For this reason, it does not make sense to use L when formulating the probabilistic model. We then preferred to stay with the notation when substituting the observations to avoid unnecessary confusion. We recognize that this distinction was not entirely consistent throughout the manuscript.

We modified the text to more clearly distinguish the terms "probability distribution of observations conditional on parameters" and "likelihood function" (of the parameters) after substituting the observations. This means we changed the wording from "likelihood" to "probabilistic model" in all the places where it does not specifically refer to inference. However, there are some exceptions; we kept the term "likelihood function" in the title and the abstract (even though we refer to the probabilistic model there), to keep the keyword and the connection to the more common terminology that does not distinguish between the two.

J. Vrugt(24): Figure 6 – the values of eta show a strong temporal correlation for error model E2 and E3. Would it be possible to plot, in some way, the decorrelated eta values (with serial correlation removed).

AC(24): What we could plot is the deviation of eta from its expected value (given the previous eta) as a function of time, which could be interpreted as decorrelated eta values.

We calculated the standardized innovations of eta for the data underlying Figure 6 and included a time series plot of those innovations for both catchments in the supplementary information.

J. Vrugt(25): In general, it may be useful if the authors include a plot of the marginal posterior distributions of the model parameters and nuisance variables. As it stands it is difficult to determine which parameters are well defined and which variables are not well defined by inference against the measured data (for one or more error models). In fact, the authors could compute the KL divergence of the prior and posterior distributions for each error model. In any case, it would be good to have insights on how well the parameters and nuisance variables are defined. Do their posterior distributions extend over the entire prior ranges, or are they limit to a small region inside the prior distribution? Note, Figure 6 goes a long way but is difficult to interpret as the matrix plot is rather small and the x-ranges are scaled according to the posterior uncertainty.

AC(25): We agree that these would be useful plots.

We included the prior and the posterior marginal density plots for all the error models, parameters and temporal resolutions for one catchment in the supplementary information. We also computed the KL-divergence and included that information in the supplementary information, in the form of a figure for both catchments, all error models and all parameters.

J. Vrugt(26): Figures 3 and 4: I find these results a bit difficult to interpret. The color/symbol coding is not necessarily clear – making it difficult to interpret the findings. I am sure the authors can find a way of plotting from which the main results are directly visible. Then, again, other readers may like to digest this plot.

AC(26): We agree that the plots are a bit crowded and can be difficult to interpret.

We made these plots more easily interpretable by enlarging the size of the individual panels, by slightly changing the color coding, adding lines for visual reference and removing the jitter.

J. Vrugt(27): Figure 5: Difficult to see the differences between the three panels. Would it be possible to enlarge the horizontal length of each of the subplots? Right now, the measured data interacts too much with the grey region, particularly when the posterior prediction/simulation uncertainty is small.

AC(27): *We enlarged the panels of Figure 5 horizontally*. *We also* adapted *the legend of Fig. 5 to include the more precise annotation of E3a* instead of E3a. In the figure caption, we mention now that this figure is based on hourly resolution. Additionally, we also enlarged the panels of Figures 6 and 9*.

J. Vrugt(28): Note, the authors use the wording "prediction" – one could argue though that what is presented are simulations as the rainfall for the next is assumed known when simulating streamflow values.

AC(28): We agree that what is input and what is predicted is a matter of systems boundaries. Thus, all predictions are conditional on some inputs. As we are dealing with hydrological and not (also) with climatological models, we still think that prediction should not lead to misunderstandings.

To clarify our system boundaries, we added a statement at the end of Section 2.2 to say that the hydrological model is evaluated for given precipitation and potential evapotranspiration data, also in the prediction period.

J. Vrugt(29): Page 24, Line 9 – 12: Is this not due in large part because of ignoring the laws of total expectation and total variance? Per my previous comment on this topic.

AC(29): As we are not ignoring the laws of total expectation and of total variance, this cannot be the reason (see our reply to comment 20). When looking at the time series of η in Fig. 9, using a constant autocorrelation time would obviously not be adequate as there are much shorter-term fluctuations during rainfall periods than during recessions. It is also clear from a hydrological point of view that (irregular) rainfall destroys the very strong autocorrelation structure we see during recession periods. The point of non-stationary autocorrelation was also raised by Th. Wöhling as referee comment 5 (Hydrol. Earth Syst. Sci. Discuss., 12, C831–C841, 2015) on the manuscript by Scharnagl et al. (2015) that was mentioned by the referee. This said, it is also clear that non-constant autocorrelation is not the only deficit of our deterministic and probabilistic models and further research is needed to further improve an adequate uncertainty description of hydrological models. However, the consideration of non-constant autocorrelation was a point that, in our view, has not been sufficiently discussed in the hydrological literature so far and we hope to contribute to stimulating this topic.

J. Vrugt(30): I think a weakness of this paper is that the authors do not compare their findings against another likelihood function. In the introduction section, the authors discuss strength and limitations of previously used/developed likelihood functions – they use this as justification for their own approach.

Yet, my own practical experience suggests that a simple AR-1 likelihood would already do quite a reasonable job. This likelihood is easy to include in the present paper. What is more, the authors should consider the generalized likelihood function – it is argued that this likelihood has a limitation because of the treatment of serial correlation on non-standardized residuals – this is easy to remedy in practice. Then, the argument of analytic tractability I do not really follow (Page 3, Line 22).

> **AC(30):** The paper does systematically compare multiple likelihood functions. They were all implemented with the same framework, to ensure comparability, but they rest on fundamentally different assumptions. For example, likelihood E2 is a "simple AR-1 likelihood". It is clearly shown in the paper that its performance is very bad in the considered case studies. We see no necessity to test another, similar version of a simple AR1 model. As for the generalized likelihood function, we agree that a comparison with the presented framework would be interesting and useful. However, since both approaches are frameworks with considerable flexibility, a meaningful comparison would require to test a large number of probabilistic models covering a reasonable range of different assumptions with both frameworks. This would go clearly beyond the scope of this study. Since we do not attempt that comparison, we do not argue that the presented framework leads to better results than the generalized likelihood function, but only repeat the concerns that have been raised by Evin et al. (2013) about the generalized likelihood. Then, we do not completely understand what the referee means by "easy to remedy in practice". It is not obvious for us how the shortcomings documented in Evin et al. (2013) could be overcome since this would require a new approach that would have to be theoretically developed and tested with a practical application. As we understand it, what comes closest to the generalized likelihood function, including corrections of the mentioned shortcomings, is the "Likelihood 2" in the submitted manuscript of Scharnagl et al. (2015). There, a heavy-tailed distribution is assumed for the innovations of the stochastic process describing the residuals, as in the generalized likelihood, but the autocorrelated process is applied to the transformed residuals, as suggested by Evin et al. (2013). However, also Scharnagl et al. (2015) obtain heavily biased results when assuming constant autocorrelation in a case where it was not appropriate to assume so. Specifically, we would suspect that the generalized likelihood function, after addressing the concerns of Evin et al. (2013), might also benefit a lot from considering non-stationary correlation, which might lead to similar results as presented in this study. This would certainly be a very interesting potential future study.
>
> *We expanded page 3, line 22 and page 5, line 10 by including more explanations about the benefits of specifying the distributional assumptions in the intuitive space of streamflow as compared to the abstract space of transformed residuals or innovations of transformed residuals*.

J. Vrugt(31): Would the inference not lead to more realistic results if the authors augment their likelihood with an error model for the rainfall data? This would carry another set of nuisance variables / hyper parameters (depending in large part on the choice of rainfall prior) but make the inference more robust.

> **AC(31):** We agree that this is another important aspect for quantifying uncertainty of hydrological models. We consider such approaches, which try to distinguish between different

sources of uncertainty explicitly, as another class of approaches that come with their own benefits and shortcomings. This study intentionally focused on an approach to describe the total uncertainty in a lumped way, which minimizes the number of error model parameters and avoids the potential identifiability problems associated with estimating input errors.

We expanded the sentence on page 2, Line 23 accordingly, mentioning the benefits and shortcomings of explicitly accounting for input uncertainty in more detail.

J. Vrugt(32): Just a thought – but is nonstationary the right wording in the present application of the likelihood function? If tau does vary between rainfall and dry periods – but these two values of tau repeat themselves in the future (e.g. are constant) – then one may argue that overall the residual time series is a stationary time series. Tau just differs between rainfall and non-rainfall days.

AC(32): We acknowledge that we chose a very simplistic non-stationary pattern. We would still call it non-stationary because of the high potential we see in relaxing the assumption of stationary autocorrelation in general, preferably also with more complex patterns.

J. Vrugt(33): Overall, I think the author should better recognize the highly related work of Scharnagl (2015) published in the same journal (HESS). Indeed, this paper used the Student distribution with the Fernandez and Steel template function for skew.

AC(33): See comment 19.

Reply to Referee #2

RC: This is an interesting well-written paper that revisits some open problems with the statistical characterization of hydrological model residuals (differences between observed and simulated values) in the context of conceptual rainfall-runoff modeling. Specifically, it addresses the issue of accounting for autocorrelation of model residuals, which is known to be troublesome in e.g. semi-arid basins where performance of spatially lumped models often is sub-optimal. The paper shows that similar problems occur in humid basins when the temporal resolution increases from daily to hourly. A novel approach that uses different autocorrelation coefficients for dry and wet periods is shown to yield better probabilistic streamflow predictions compared to the common practice of using a constant autocorrelation coefficient.

> **AC:** Thank you for this general feedback.

RC (1): Title and contribution: the title is quite broadly formulated and doesn't really bring out the main novel contribution of the paper, i.e. improved autocorrelation modeling at sub-daily resolutions. In my opinion the proposed likelihood function framework is secondary to this: although it is different from previous approaches, its performance for constant autocorrelation is similar to previous approaches (at least qualitatively – a numerical comparison is not done in the paper), and the novel use of a variable autocorrelation coefficient could also readily be implemented with previous approaches. So it's not entirely clear what we gain from the new framework, even though I do find it quite elegant. If the main selling point is the new likelihood framework then more extensive comparisons (both theoretical and empirical) with existing approaches would be helpful. The proposed framework also has some (conceptual) issues, as discussed in the next point.

> **AC (1)**: The referee correctly points out that the two major elements of the manuscript, the likelihood framework and the variable autocorrelation coefficient are independent of each other. While we do want to stress that the latter can strongly improve the results of the inference procedure, we do not claim that the presented likelihood framework leads to better results than other approaches. For this it is too general, the results achieved with the framework will depend strongly on the assumptions made. It will indeed lead to similar results as previous likelihoods, if the assumptions made are very similar (e.g. constant correlation where it is not appropriate). The major novelty of the framework is the ability to transparently discuss the assumptions about the distribution of streamflow given the model output, as the next comment of the referee illustrates. With previously used approaches like Box-Cox transformations or the generalized likelihood, the assumed distribution of streamflow is often unknown and cannot be efficiently communicated and discussed. We do acknowledge that this benefit is of rather qualitative nature and cannot be illustrated by a quantitative comparison.
>
> *We included some more theoretical explanations about the potential benefits of the likelihood framework.*
>
> We also agree that the variable autocorrelation coefficient is among the most important novel contributions of the paper.

Therefore, we included it in the title of the paper, which was changed to: "A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation"

RC(2): Section 2.1: the statistical model and corresponding likelihood is based on specifying the density of observed discharge Q conditioned on simulated discharge, Eq. 1. To avoid negative Q values, the density is truncated at zero by removing all probability mass for Q<0 and placing it at Q=0. This deviates from the usual truncation approach, which would scale the entire density by 1/(1-FQ(0)). In fact, the proposed approach results in strange bimodal looking densities with a peak at Q=0 and another at some Q>0; somehow I don't think this is an intuitive model that hydrologists would come up with based on prior knowledge (as suggested on page 5, line 11)! Another consequence of the chosen truncation is that the transformed variables eta in Eq. 2 are also truncated and not Gaussian. This is partially acknowledged on page 5 line 28, but I don't think it's correct that the lower tail of eta will be lighter: there simply will be no lower tail (truncation). Note that these issues could be remedied by adopting the usual truncation approach (scale the entire density) or by using a density with nonnegative support. It's not clear whether these truncation issues matter in practice, perhaps not for the humid basins studied here, but it may matter in drier basins with discharge close to zero.

AC(2): This critique is partly based on a misunderstanding resulting from an insufficient discussion in our paper. The intention of our approach was to allow for a finite probability at Q=0 which is important for intermittent rivers and is often poorly reflected by the deterministic part of the model in which the discharge approaches zero only asymptotically. Such a finite probability for Q=0 can be desirable, see e.g. Smith et al. (2010), whose approach of a mixture distribution is in conceptual agreement with Eq.(1). However, truncation and assignment of the truncated mass to Q=0 is only needed if the distribution extends to negative values. As our framework allows for an arbitrary distribution of discharge, we can choose the distribution mentioned by the referee from the beginning (truncate at zero and rescale the density accordingly). In this case there will not be any additional truncation in our probabilistic model and the probability of Q=0 will be zero. We agree with the referee that this may often be the choice of the modeler and this option is fully covered by our framework (which we did not clearly write in the paper so far). In our study we chose the other distribution to illustrate the possibility of having a finite probability for Q=0, as suggested by Smith et al. (2010). This probability distribution can look a bit non-intuitive, but we still believe that e.g. hydrologists working in ephemeral catchments would appreciate having a finite probability for Q=0.

In Section 2.1, we improved the discussion of the flexibility the modeler has in choosing the distribution of discharge, in particular regarding distributing the probability for a negative outcome to all positive values of discharge. We also added a statement concerning the limitation of the truncation and rescaling approach on page 2, Line 9. To clearly distinguish between the two different "truncating" approaches, we reserved the word "truncating" for the approach that includes rescaling of the positive part of the distribution. Thus, the approach presented in the case studies, where the probability of a negative outcome is assigned to zero, is not called "truncation" anymore.

Concerning the truncation of the lower tail, the referee is correct in the statement that there is no lower tail at all anymore for each individual time step. However, when considering the marginal distribution of the etas at all the time steps, there is still a lower tail, since the individual distributions at each time step are each truncated to a different extent at the lower end, which still results in a continuous marginal distribution over all time steps. We admit that this is not mentioned clearly enough in the manuscript.

We complemented page 5, line 28 by mentioning that we mean the marginal distributions of the etas, which still has a lower tail.

RC(3): Section 2.3, evaluation criteria: the reliability and precision metrics are counterintuitive in that smaller values for these metrics indicate better performance. Unreliability and imprecision metrics? Another natural metric to consider is the maximum loglikelihood value of each model (perhaps corrected with number of parameters, as in BIC).

AC(3): In order to maintain consistency with McInerney et al. (2017), we would like to keep the names "reliability" and "precision". However, we agree that the names can be misleading in this case.

We added 2 arrows in Figure 3, clarifying that smaller reliability and precision values mean better results. We also added corresponding sentences in the captions of Table B1 and B2.

We agree that the maximum loglikelihood value would be another straightforward metric to consider, but is not clear what information we would gain from it that we do not already have in the other measures. It is not very meaningful for practical purposes and it says nothing about the quality of the predictive distribution since it only characterizes the single best model realization. Given also the broad range of measures already included in the study, we would prefer to not include the maximal loglikelihood as a measure.

RC(4): Section 3.3, error models: the method of Fernandez and Steel (1998) to skew a symmetric density was also used by Schoups and Vrugt (2010), in their case to skew an exponential-power density. It may be appropriate to cite that paper here, especially if that's where you learned about the Fernandez and Steel method.

AC(4): Thank you for pointing this out. We were not aware that Schoups and Vrugt (2010) have already used the approach of Fernandez and Steel (1998). We found it independently in the statistical literature.

We mentioned that already Schoups and Vrugt (2010) have used this skewing approach.

RC(5): Table 2, page 14, line 1: the E1 model also truncates fQ at zero, which is another difference with maximizing NSE.

AC(5): This is right, the two approaches are different in their assumption about the distribution of streamflow in the range of Q<=0. In any case where Q<0 is not observed (almost always), the negative part of the distribution will not affect inference, and therefore it will not lead to a

different result. However, when there is data of Q=0, the NSE maximization might lead to different results than E1.

We added this statement on page 14, line 1.

RC(6): Figure 6: in the top-right plot for model E3, it's not clear that distributional assumptions for eta are satisfied; there are significant outliers in this plot, and the variance is not constant.

AC(6): We agree with the statement of the referee. We do not claim that error model E3 results in perfectly fulfilled assumptions. Also the assumption of zero correlation during precipitation events is violated. We think that the major benefit of E3 is a pragmatic trade-off between fulfilling the assumptions to a satisfactory degree and still providing reasonable results in terms of the fit of the hydrological model and the predictive uncertainty.

RC(7): Conclusions: finding 5 (accounting for autocorrelation is good) seems to contradict finding 1 (accounting for autocorrelation can be bad); you may need to clarify/reformulate these a bit.

AC(7): We agree that there is some contradiction in those two findings as they are formulated now.

We added a statement in finding 5 saying that the benefit of considering autocorrelation is only useful if the problems mentioned in finding 1 can be avoided.

RC(8): Conclusions: finding 3 states that errors in streamflow are expected to be less correlated during precipitation events than during dry weather. Is that always the case though? What about rainfall errors, these could lead to significant bias and correlated errors in simulated streamflow. Also, structural errors in the fast flow component of the model may be (much) larger than in the slow flow component. Perhaps a better, more general, justification for a nonstationary correlation model is to say that the error correlation structure can be expected to differ between wet and dry periods (for various reasons), and then let the data decide whether wet or dry has the larger autocorrelation coefficient.

AC(8): We agree with that statement. It is not completely certain a-priori that the rainfall events will have the less correlated errors, although we do believe that this will often be the case.

We added a statement on page 24 line 16 and adapted finding 3 of the conclusions to say that in principle also higher correlation of errors during precipitation events is possible, *but we still mention that reduced correlation is more likely.*

RC(Edits):

- page 4, line 5, "Understanding...remains poorly understood": remove "understanding".

- Eq. 13, Nash-Sutcliffe formula: change Q to Qobs in the denominator

- page 21, line 12: "normality" has a typo

- figure 9, caption: left/right should be top/bottom

- page 27, line 28: likeli -> likely

- page 28, line 17: "appropriate" has a typo

AC(Edits): Thank you for pointing these out. We agree that all of these are errors.

We corrected them in this version of the manuscript.

Additional Changes

Page 1, Line 23: Highlighted the special characteristic of observation errors: they only need to be accounted for when doing inference based on observations.

Page 2, Line 16: Based on a comment of Alberto Montanari, the non-negativity is not listed as a separate characteristic anymore, but added to the first point of the list about the non-normality of the residuals.

Page 2, Line 26: More precise wording regarding the treatment of heteroscedasticity in the weighted least squares approach.

Page 2, Line 29: Bibtex Entry was corrected: Del Giudice, D.

Page 4, Line 11: some clarifications w.r.t. the disadvantages of not accounting for autocorrelation in the residuals

Page 4, Line 21 and Line 30: some clarifications regarding the goals of this study

Page 5, Line 14: better wording regarding the application of copulas to access DQ by Wani et al.

Page 8, Line 16: more precise wording regarding the criteria of performance of an error model.

Page 8, Eq. (10): the absolute value operator was missing in the numerator.

Page 10, Eq. (14): replaced the integration with a sum

Page 13, Figure 2: We included the reference modelled streamflow, Qdet, and changed the centering of DQ so that Qdet is equal to the mean of DQ and not the mode. This is more representative of the method applied for generating the results in the paper.

Page 13, Line 15: Since Qobs is a deterministic value, the notation E[] does not make sense, we replaced it with the “bar” to denote the average.

Page 14, Line 12: Short sentence clarifying the need to account for temporal lags between precipitation and streamflow when using E3.

Additional references

Hernández-López, M. R. and Francés, F.: Bayesian joint inference of hydrological and generalized error models with the enforcement of Total Laws, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-9, 2017.

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.

A likelihood framework for ~~likelihood functions of~~ 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)

Abstract. The widespread application of deterministic hydrological models in research and ~~practise~~ 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, ~~the~~ residual error models ~~used to describe them~~ 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)~~, and~~ . When performing inference, observation errors (iv) ~~observation errors of~~

~~the output,~~ are an additional source of uncertainty, which arises for example due to errors in rating curves (e.g. Kuczera and Franks, 2002). The sources (i-iv) usually result in residual errors of ~~streamflow predictions~~ 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 ~~, and charactered~~ 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. ~~For example~~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.

- ~~Non-negativity: in typical situations, streamflow measurements are non-negative meaning that streamflow always flows in the same direction~~

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 ~~requires information about input and output uncertainty to overcome an otherwise ill-posed problem, and is not pursued in this work~~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:

- ~~Residuals are often normalised for weighed~~ 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 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 ~~likelihood function~~ 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 (e.g. Kloeden and Platen, 1995; Yang et al., 2007).

- Non-negativity of streamflow can be addressed by truncating the error pdf so that it does not extend to negative streamflow. ~~However, this~~ 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. 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 ~~likelihood parameterisation~~ 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 ~~formulating~~ describing the errors with an autoregressive process with a skew exponential power (SEP) rather than a normal distribution ~~. This results in marginal distributions for streamflow at given time points that can be non-normal and skewed, but are not easily accessible analytically. Furthermore, the approach was~~ 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 standardized 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 the ones 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 transformation 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) ~~compared~~ 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 ~~on~~ 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) ~~recommended inferring autoregressive parameters~~ 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. ~~Understanding the~~ The conditions under which this can be achieved ~~remains~~ remain poorly understood.

Fourth, the potential advantages of more complex error models are under-appreciated by the hydrological community. ~~Most commonly, residual error models are used to plot some "uncertainty bands " around the hydrograph. For such purposes~~For relatively simple uncertainty analysis, like the plotting of uncertainty bands around hydrographs, the use of ~~relatively~~ 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 ~~unrealistic values of hydrograph signatures , particularly if these~~ too strong short term fluctuations, which have a large effect on hydrograph signatures that are sensitive to noise, such as ~~for example~~ the Flashiness Index (Baker et al., 2004; Fenicia et al., 2018). The ability of correctly representing 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 in wet and dry periods, unequally spaced observation time points and non-negativity of streamflow.

2. ~~Investigate the ability to infer the various parameters of the error model and the~~ 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?

~~Note that the developed framework allows for additional flexibility in aspects that are not covered with Questions 2a-2e (e.g. unequally spaced observations, non-negativity). To limit the scope of this paper, we refrain from controlled experiments w.r.t. those aspects.~~ The paper is structured as follows. The theoretical framework for ~~likelihood functions~~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 setup 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 ~~Likelihood~~ 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 ~~likelihood function~~ probabilistic framework for hydrological models, since ~~many modellerswill have an intuitive idea about the probability distribution of the observations given an output of their model (Fig. 1).~~ 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 (Bates and Campbell, 2001; Del Giudice et al., 2013; McInerney et al., 2017, e.g.) 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 ~~. Note the major difference to the approach of Schoups and Vrugt (2010) and transformation based approaches (Bates and Campbell, 2001; Del Giudice et al., 2013; McInerney et al., 2017, e.g.), where D_Q is not easily accessible.~~ in units they are familiar with. Wani et al. (In preparation) present another approach in which ~~the temporal dependence of~~ 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. ~~$Q(t) \sim D_Q(Q_{\text{det}}(t), \boldsymbol{\psi})$~~ $Q(t) \sim D_Q(Q_{\text{det}}(t, \boldsymbol{\theta}), \boldsymbol{\psi})$, 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 standard deviation), but these distributions belong to the same parametric family ~~D_Q~~. 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

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

where f_{D_Q} and F_{D_Q} ~~be~~ are the density and cumulative distribution function of D_Q, respectively. p is a probability density for $Q > 0$ and a discrete probability for $Q = 0$. Note that Eq. (1) reflects our prior knowledge that $Q \geq 0$ ~~. Transforming Q according to~~ 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. (~~2) leads to a (potentially correlated) time series: with~~ 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. Eq. (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 ~~Q_{det} is the result of the deterministic hydrological model~~ 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, ~~η will be marginally standard normally distributed,~~ except for the ~~truncation at zero~~lower tail, which can ~~lead to lighter tails on the lower end.~~

be lighter due to the truncation at zero at each individual time step. If there are systematic deviations of Q from Q_{det}, the series $\eta(t_i)$ will be autocorrelated.

To describe autocorrelation ~~between successive streamflow values~~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:

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

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 ~~$\tau(t_{i+1})$ represents the~~ a characteristic correlation time, $\tau(t_i)$, that is assumed to be constant over the interval ~~$[t_i, t_{i+1}]$.~~ ~~Consider the conventional AR(1) process: where ϵ_t is white noise. For equidistant time steps, the relation between τ in Eq. (4) and ϕ in Eq. (5) is given by: where Δt is the size of the time step. To apply Eq.~~ $[t_{i-1}, t_i]$.

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

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

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

with

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

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~~. One advantage of this formulation is that it ~~allows us~~ 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 ~~likelihood~~ probability of the streamflow Q, we ~~derived~~ used Eq. (1) to (4) to derive the following conditional probabilities for ~~$Q(t_{i+1})$ given $Q(t_i)$~~ $Q(t_i)$ given $Q(t_{i-1})$ (see Appendix A for the full derivation):

if $Q(t_{i-1}) > 0$:

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

if $Q(t_{i-1}) = 0$:

$$p_i\big(Q(t_i) \mid Q(t_{i-1}), \boldsymbol{\theta}, \boldsymbol{\psi}\big) = \begin{cases} f_{D_Q\left(Q_{\mathrm{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\right)}\big(Q(t_i)\big) & \text{for } Q(t_i) > 0 \\ F_{D_Q\left(Q_{\mathrm{det}}(t_i,\boldsymbol{\theta}),\boldsymbol{\psi}\right)}(0) & \text{for } Q(t_i) = 0 \end{cases}$$

Note that p is a probability density (denoted by f) if ~~$Q(t_{i+1}) > 0$~~ $Q(t_i) > 0$, and an integrated, discrete probability (denoted by F) if ~~$Q(t_{i+1}) = 0$. The full~~ $Q(t_i) = 0$. 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 $\tau \to 0$, i.e. if correlation is absent or neglected.

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

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

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.

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 df, 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 $df \to \infty$. Two assumptions are tested to centre D_Q at Q_{det}:

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

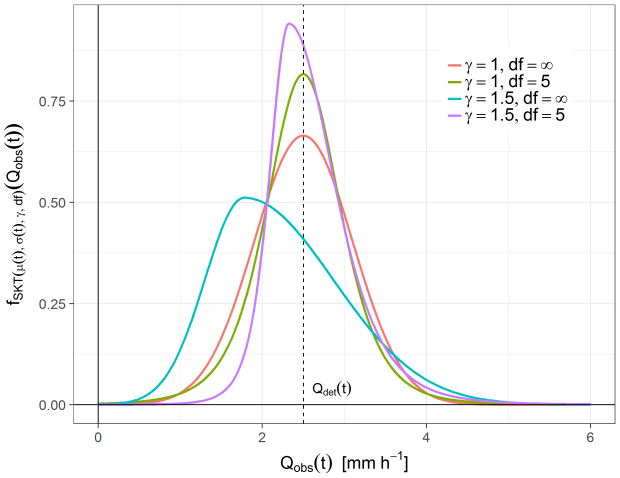

Figure 1. Example of skewed Student's t-distributions with $E[D_Q] = Q_{\mathrm{det}}(t) = 2.5$ mm h^{-1} and standard deviation $\sigma_{D_Q}(t) = 0.6$ for different values of skewness, γ, and kurtosis, df.

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

i.e. we either assign the expected value or the highest probability 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:

$$\sigma_{D_Q}(t) = aQ_0 \left(\frac{Q_{\mathrm{det}}(t)}{Q_0} \right)^c + bQ_0 \tag{10}$$

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 , we set $c = 1$. 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 lists the error models applied in this study, together with their underlying assumptions. E1 is included as a reference

Table 1. Overview of the error models applied in this study and their corresponding parameters (×: fitted). If * is appended to the name of the error model, a smoothed version of $P_{\mathrm{err}}(t)$ (moving average of window size 5 h) was used in Eq. (11).

Error Model	$\tau_{\min}$	$\tau_{\max}$	a	b	γ	df
E1	0	0	×	×	1	∞
E2	$=\tau_{\max}$	×	×	×	1	∞
E3$^{(\star)}$	0	×	×	×	1	∞
E3a$^{(\star)}$	×	×	×	×	1	∞
E4$^{(\star)}$	0	×	×	×	×	×
E4a$^{(\star)}$	×	×	×	×	×	×

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 $Q_{\mathrm{obs}} = 0$, are identical to the ones made when e.g. maximising the Nash-Sutcliffe Efficiency, or, equivalently, minimising the squared residuals. Error Model E2 represents a conventional approach of considering autocorrelation. In the case of equally spaced time-steps, it is similar to the error model applied e.g. by Evin et al. (2013), 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 $Q_{\mathrm{obs}} = 0$. 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:

$$\tau(t) = \begin{cases} \tau_{\min} & \text{for } P_{\mathrm{err}}(t) > 0 \\ \tau_{\max} & \text{else} \end{cases} \tag{11}$$

where P_{err} is the precipitation used as an input for the error model. In E3, $\tau_{\min}$ is fixed at 0, while in E3a, it is fitted. P_{err} was either equal to the recorded precipitation, P, or, in 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 $\tau_{\min}$ and $\tau_{\max}$ during precipitation events, and to be more robust w.r.t. 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, $P_{\mathrm{err}} = P$ 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 $\tau_{\min}$ 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, $\boldsymbol{Q}_{\mathrm{det}}$ and $\boldsymbol{Q}_{\mathrm{obs}}$. 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, ~~Q_{det}~~ $\boldsymbol{Q}_{\mathrm{det}}$, for all time points.
2. Using $\boldsymbol{\psi}$ and ~~Q_{det}~~ $\boldsymbol{Q}_{\mathrm{det}}$, calculate the likelihood in Eq. (8)~~by substituting the argument Q with the observed streamflow, Q_{obs}~~.

As the likelihood (Eq. 8) ~~can efficiently be evaluated analytically~~ 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):

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

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 ~~,~~ $\boldsymbol{Q}_{\mathrm{det},j}$ ~~,~~ 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 ~~$Q_{\mathrm{det},j}$ and~~ $\boldsymbol{\psi}_j$ and $\boldsymbol{Q}_{\mathrm{det},j}$, 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 (~~Appendix ??~~see supplementary material).

2.4 Evaluation criteria

How can the performance of empirical error models, as the ones 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 following criteria: (a) good reproduction of observed dynamic fluctuations by individual model realizations, (b) good overall predictive ~~distributions and~~ marginal distribution of streamflow (c) small absolute deviance between model output and observations. The Flashiness Index (Sect. 2.4.1) ~~quantifies~~ is an indicator for (a). The reliability and the precision 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) ~~, I_F,~~ is given by:

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

where ~~$\boldsymbol{Q} = (Q(t_1), Q(t_2), \ldots, Q(t_N))$. I_F~~ is $\boldsymbol{Q} = (Q(t_0), Q(t_1), \ldots, Q(t_n))$. 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, ~~Q_{obs}~~ $I_{\mathrm{F,obs}} = I(\boldsymbol{Q}_{\mathrm{obs}})$, the output of the deterministic hydrological model~~at the maximum posterior parameter values, $\widehat{Q}_{\mathrm{det}}$~~, $\widehat{I}_{\mathrm{F,det}} = I(\widehat{\boldsymbol{Q}}_{\mathrm{det}})$, and the individual stochastic realisations of the ~~full predictive distribution of streamflow , Q_j. The resulting metrics are denoted as $I_{F,\mathrm{obs}}$, $\widehat{I}_{\mathrm{F,det}}$ and I_F, respectively, where the latter is the median of the flashiness indices of the individual realisations Q_j. I_F~~ is predictive streamflow sample, $I_{\mathrm{F}} = \mathrm{median}(I(\boldsymbol{Q}_j))$. I_{F} is sensitive to the amount of autocorrelation in a streamflow time series, as well as the height of the peaks of ~~Q_{det}~~ $\boldsymbol{Q}_{\mathrm{det}}$ (since $\boldsymbol{Q}_j$ depends on $\boldsymbol{Q}_{\mathrm{det}}$).

2.4.2 Reliability

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

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

where ~~$\Psi = \{F_{Q(t_i)}(Q_{\mathrm{obs}}(t_i)) | i \in \mathbb{N}, i \leq N_t\}$~~ $\Psi = \{F_{Q(t_i)}(Q_{\mathrm{obs}}(t_i)) | i \in \mathbb{N}, 0 \leq i \leq n\}$, F_{Ψ} is the empirical cumulative distribution function of Ψ 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 smaller values of Ξ_{reli} correspond to better, and zero to perfect, reliability. It summarises the deviance of the observations from the predictive distribution over all time points, and the distance is measured in the uniform space. Therefore, the influence of heavy outliers on Ξ_{reli} is limited.

2.4.3 Precision

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

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

where ~~$\sigma_{Q(t_i)}$~~ $\sigma_Q(t_i)$ 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. $\Omega_{\mathrm{prec}} \in \mathbb{R}^+$, and small values of Ω_{prec} indicate high precision or small predictive uncertainty. The smaller the predictive uncertainty, the better the quality of the underlying model, given that the predictions are not overconfident.

2.4.4 Nash-Sutcliffe Efficiency

The Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970), ~~E_N~~ $E_{\mathrm{N,f}}$ (f for function), is defined as:

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

where ~~$\boldsymbol{Q} = (Q(t_1), Q(t_2), \ldots, Q(t_N))$~~ $\boldsymbol{Q} = (Q(t_0), Q(t_1), \ldots, Q(t_n))$. It is used in this study to ~~quantify the agreement between $\widehat{Q}_{\mathrm{det}}$ and Q_{obs},~~ assess the output of the hydrological at the maximum posterior parameter density, $\widehat{E}_{\mathrm{N,det}} = E_{\mathrm{N,f}}(\widehat{\boldsymbol{Q}}_{\mathrm{det}}, \boldsymbol{Q}_{\mathrm{obs}})$, as well as ~~between the j-th stochastic realisation Q_j and Q_{obs}. The two cases are denoted as $\widehat{E}_{\mathrm{N,det}}$ and E_N, respectively, where E_N is the median of the efficiencies of the individual realisations Q_j~~ the stochastic simulations, $E_{\mathrm{N}} = \mathrm{median}(E_{\mathrm{N,f}}(\boldsymbol{Q}_j, \boldsymbol{Q}_{\mathrm{obs}}))$. 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:

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

It is calculated w.r.t. the ~~maximum posterior output of the deterministic model; $\widehat{\Delta}_{\mathrm{Q,det}} = \Delta_Q(\widehat{Q}_{\mathrm{det}}, Q_{\mathrm{obs}})$~~ model output based on the parameter values at the maximum posterior density; $\widehat{\Delta}_{\mathrm{Q,det}} = \Delta(\widehat{\boldsymbol{Q}}_{\mathrm{det}}, \boldsymbol{Q}_{\mathrm{obs}})$, as well as for the ensemble of individual stochastic simulations: ~~$\Delta_Q = \mathrm{median}(\Delta(Q_j, Q_{\mathrm{obs}}))$~~ $\Delta_Q = \mathrm{median}(\Delta(\boldsymbol{Q}_j, \boldsymbol{Q}_{\mathrm{obs}}))$. 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 ~~averaged hydrographs~~ average hydrograph.

3 Case study setup

3.1 Catchments and data

The ~~likelihood~~ probabilistic framework developed in Sect. 2.1 was tested in two case study sites, the Murg and the Maimai ~~catchmets~~catchments, which are described in this section. The Murg river flows through a hilly headwater catchment in 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 part 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 consists of hourly averages of streamflow, precipitation and potential evapotranspiration from January 1995 to December 2002. Calibration was performed in the first 5 years (Jan 1995-Dec 1999) and validation in the consecutive 3 years (Jan 2000-Dec 2002). Streamflow data is a courtesy of the Swiss Federal Office for the Environment (FOEN). Precipitation and potential evapotranspiration are based on meteorological data (Meteoschweiz, 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 Jan 1985-Dec 1986, and validation during Jan-Dec 1987. The data was 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 setup 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~~ R_C the runoff coefficient (calculated from cumulative streamflow and precipitation). $Q_{\mathrm{obs,max}}$, $Q_{\mathrm{obs,min}}$ and $\overline{Q}_{\mathrm{obs}}$ are the minimum, the maximum and the average streamflow, respectively. $I_{\mathrm{F,obs}}$ is the Flashiness Index (Baker et al., 2004).

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

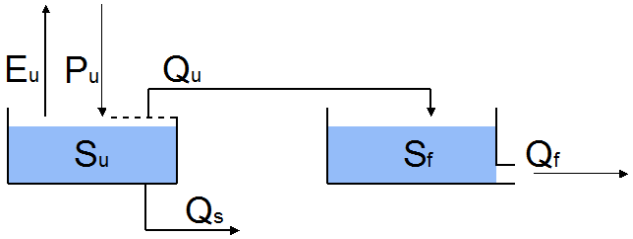

Figure 2. Structure of the deterministic hydrological model used in this study. P_u is the precipitation and E_u the evapotranspiration. S_u represents the active water content of the unsaturated zone, while S_f is a non-linear reservoir representing the fast flow component.

3.2 Deterministic Hydrological Model

The hydrological model used throughout this study is a simple, lumped bucket model with two ~~elements~~ reservoirs (Figure 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 ~~$Q_{\mathrm{det}}(t,\boldsymbol{\theta}) = Q_s + Q_f$, where $\boldsymbol{\theta} = (C_e, S_{\mathrm{max}}, k_u, k_f)$ are the~~ $Q_{\mathrm{det}}(t,\boldsymbol{\theta}) = Q_{\mathrm{s}}(t,\boldsymbol{\theta}) + Q_{\mathrm{f}}(t,\boldsymbol{\theta})$, 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~~ E_u, P_u, Q_u, Q_s, Q_f) and states (~~S_u, S_f~~ S_u, S_f) of the model are given by:

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

$$\begin{aligned}
\frac{dS_f}{dt} &= Q_u - Q_f \\
Q_f &= k_f {S_f}^{\alpha}
\end{aligned} \tag{19}$$

where ~~E_p~~ E_p is the potential evapotranspiration. ~~The following parameters were kept fixed: $m = 0.01$, $\beta = 3$, and~~ 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 $\alpha = 2$ ~~.~~

3.3 ~~Error Models~~

~~For D_Q, we use the skewed Student's t-distribution (Fig. 1) as the most general case, which is obtained by transforming the conventional Student's t-distribution according to Fernandez and Steel (1998). Thus, we introduce two error model parameters: γ, defining the degree of skewness, and df, the degrees of freedom as a measure for the kurtosis. The skewed Student's t-distribution reduces to the normal distribution for $\gamma = 1$~~ were found to lead to reasonable results in both investigated catchments and were fixed due to potential interactions with S_{max} and ~~$df \to \infty$. Two assumptions are tested to centre D_Q at Q_{det}: i.e. we either assign the expected value or the highest probability of D_Q to Q_{det}. A third alternative would be to set the median of D_Q equal to Q_{det}. It is a priori unclear which of those options is most suitable. 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.~~ 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.

~~Shape of the skewed Student's t-distribution for different values of skewness, γ, and kurtosis, df.~~

~~The standard deviation of D_Q is parameterised as follows: 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. Like previous studies (Thyer et al., 2009; Schoups and Vrugt, 2010; Evin et al., 2013), we set $c = 1$ in Eq. (10). 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, which focuses on the autocorrelation of the errors, we set $c = 1$. This also leads to better comparability to the aforementioned studies. Note that the parameters a and b become dimensionless (and therefore more universal) by including a reference streamflow, $Q_0 = E[Q_{\mathrm{obs}}]$. Thus, a can be seen as the standard deviation of the error relative to the modelled streamflow, while b represents a characteristic fraction of the reference streamflow Q_0, below which the magnitude of the error starts to become less dependent on Q_{det}.Overview of the error models applied in this study and their corresponding parameters ($\times$: fitted). If * is appended to the name of the error model, a smoothed version of $P_{\mathrm{err}}(t)$ (moving average of window size 5 h) was used in Eq. (11). Error Model $\tau_{\min}$ $\tau_{\max}$ a b γ dfE1 0 0 $\times$ $\times$1∞E2 = $\tau_{\max}$ $\times$ $\times$ $\times$1∞E3$^{(\star)}$ 0 $\times$ $\times$ $\times$1∞E3a$^{(\star)}$ $\times$ $\times$ $\times$ $\times$1∞E4$^{(\star)}$ 0 $\times$ $\times$ $\times$ $\times$ $\times$E4a$^{(\star)}$ $\times$ $\times$ $\times$ $\times$ $\times$ $\times$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, are identical to the ones made when e.g. maximising the Nash-Sutcliffe Efficiency, or, equivalently, minimising the squared residuals. Error Model E2 represents a conventional approach of considering autocorrelation. In the case of equally spaced time-steps, it is similar to the error model applied e.g. by Evin et al. (2013), who assume that the rescaled errors follow an AR(1) process with a standard normal marginal distribution. One difference between the two approaches is that we truncate D_Q at zero. 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:

where P_{err} is the precipitation used as an input for the error model . In E3, $\tau_{\min}$ is fixed at 0, while in E3a, it is fitted. P_{err} was either equal to the recorded precipitation, P, or, in 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 $\tau_{\min}$ and $\tau_{\max}$ during precipitation events, and to be more robust w.r.t. potential time lags between observed precipitation and streamflow. Since in the Murg catchment, smoothing did not change the results substantially, $P_{\mathrm{err}} = P$ 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 $\tau_{\min}$ is inferred.The prior distributions of all the parameters , listed in Table 3, were

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 believe than e.g. 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.

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	low. bound.	up. bound.
~~C_E~~ C_e	~~N~~ N	-	1	0.2	0.2	3
$S_{\max}$	~~LN~~ LN	mm	148	1086	2.7	1086
k_u	~~LN~~ LN	h^{-1}	1.8e-2	0.13	2.3e-6	5e-2
k_f	~~LN~~ LN	h^{-1}	0.37	2.7	2.3e-6	0.37
a	~~LN~~ LN	-	0.2	0.2	-	-
b	~~LN~~ LN	-	0.1	0.1	1e-2	0.5
$\tau_{\max}$	~~LN~~ LN	h	148	1086	0	2000
γ	~~LN~~ LN	-	1	0.2	0.1	5
df	~~LN~~ LN	-	14	17	3	-

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 precision 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 error in cumulative streamflow, Δ_Q, and about $\widehat{E}_{\mathrm{N,det}}$. 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 (Fig. 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 $\widehat{E}_{\mathrm{N,det}}$ (Fig. 4): for 6-hourly and daily resolution of the data, the worst performing error model in the Maimai catchment has a better $\widehat{E}_{\mathrm{N,det}}$ 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 case of high temporal resolutions in both catchments (Fig. 3, the difference between the observed and the median of the predicted Flashiness Index is around -0.4 for both catchments). ~~The~~

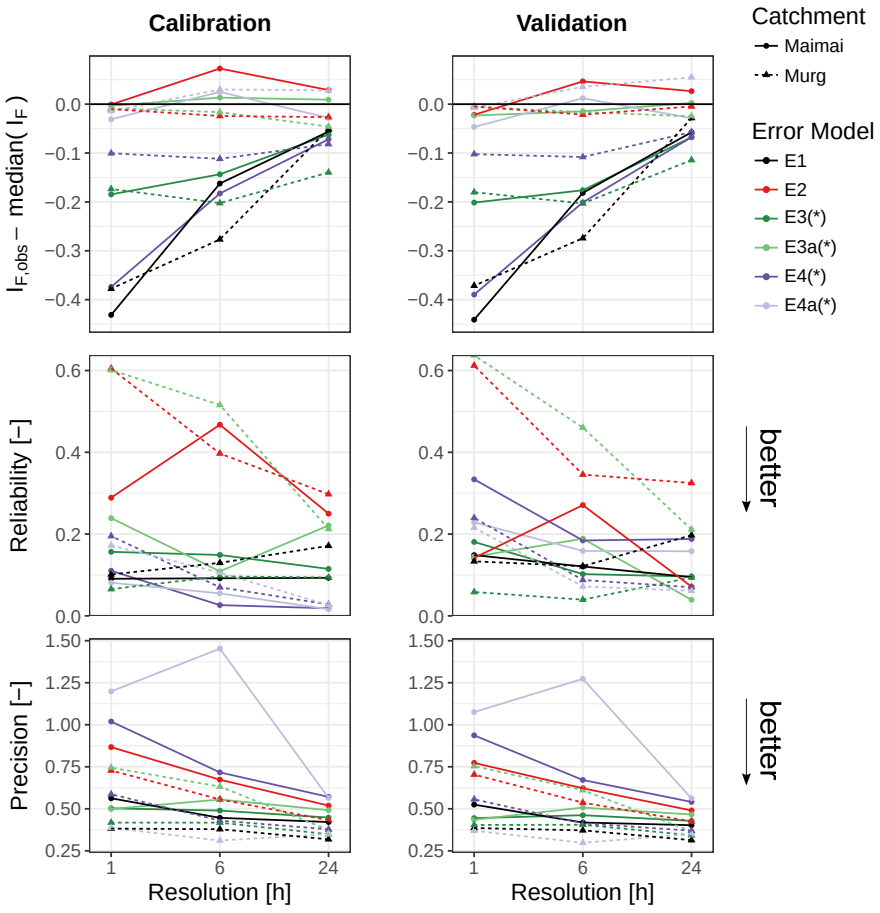

Figure 3. Performance of the error models w.r.t. flashiness index, reliability and precision for both catchments and all temporal resolutions. P_{err} was smoothed (*) exclusively for hourly data in the Maimai catchment.

~~reliability of~~ In terms of reliability, E1 is never the single best of the error models, but always among the best, and it is robust in light of varying temporal resolution (Ξ_{reli} is smaller or equal to 0.2 in all the cases, Fig. 3). E1 is also among the error models that provide the most precise predictions (average of 0.41 over all the cases), have the smallest Δ_Q (usually between 0 and -10 %) and the highest $\widehat{E}_{\mathrm{N,det}}$ overall (Fig. 4). Except for the Flashiness Index, its performance stays stable for ~~high temporal resolutions of the~~ high-frequency data in both catchments. However, the high Flashiness Index of this model demonstrates the strong violation in the description of the output behavior despite its good performance regarding the other performance metrics.

4.1.2 Model E2

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

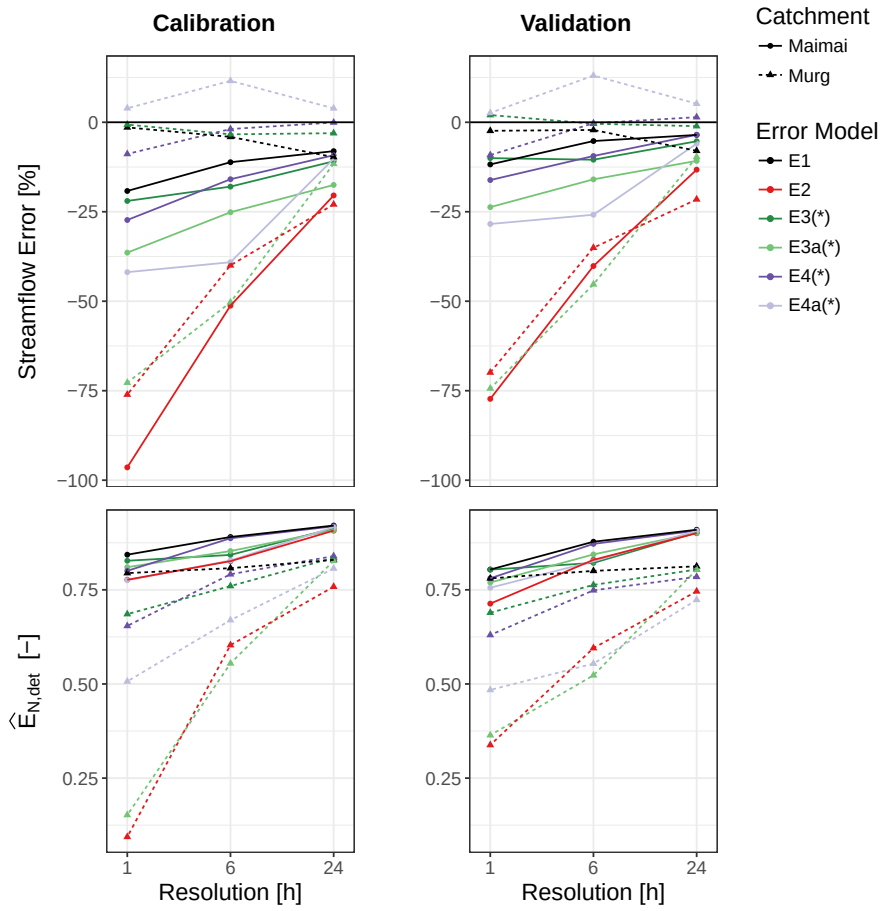

Figure 4. Performance of the error models in terms of the relative cumulative error in streamflow, Δ_Q, and the Nash-Sutcliffe Efficiency, $\widehat{E}_{\mathrm{N,det}}$, for both catchments and all temporal resolutions. P_{err} was smoothed (*) exclusively for hourly data in the Maimai catchment.

of the model 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. 3). E2 has an average precision of 0.61 over all the cases, compared to a precision of 0.41 of E1. It tends to produce large errors in cumulative streamflow, especially in case of hourly resolution ($\Delta_Q < -75\%$, Fig. 4). The degradation of the streamflow error and $\widehat{E}_{\mathrm{N,det}}$ with increasing ~~temporal resolution~~ measurement frequency is very pronounced for E2 compared to the other error models (Fig. 4).

4.1.3 Model E3

E3 generally overestimates the true Flashiness, i.e. ~~I_F~~ 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. 3). 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 smaller than 0.2 in every case and smaller than 0.1 in more than half of the cases (Fig. 3). In the validation period in the Murg catchment, it is the most reliable error model of all. The precision 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. 3). The absolute value of Δ_Q is never larger than 25 % and usually smaller than 10 % (Fig. 4). 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 ~~temporal resolution~~ 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~~ I_{F} and $I_{\mathrm{F,obs}}$ in all cases (Fig. 3, the deviation is never larger than 0.05). In the Maimai catchment, the reliability measure shows a stable performance in, with values between 0.04 and 0.19 in the validation period (Maimai, Fig. 3), showing no clear signs of worse performance for ~~high temporal resolutions~~high-frequency data. The inferred values of τ_{min} were in 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 degenerating performance of E3a with increasing ~~temporal resolution~~measurement frequency, with values of $\Xi_{\mathrm{reli}} > 0.5$ for 6-hourly and hourly data (Fig. 4), indicating poor performance. All the other metrics show a similar pattern (Fig. 4). The inferred τ_{min} were between 50 and 100 h, where values on the upper end of the spectrum coincided with bad ~~reliabilies~~ reliabilities (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~~ I_{F} is usually between -0.2 and -0.1 (Fig. 3). ~~I_F~~ 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 ~~variability~~ 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. 4). 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~~ I_{F} and $\widehat{I}_{\mathrm{F,det}}$ is smaller (Table B2). Concerning the reliability, Ξ_{reli} is largely smaller than 0.2, indicating well-conditioned predictive distributions, except in the validation period for hourly resolution (Fig. 3). In the Maimai catchment, reliability is better in the calibration period compared to the validation period, which is a sign of over-fitting. Especially for daily resolution, E4 provides very good reliabilities in the calibration period ($\Xi_{\mathrm{reli}} < 0.03$, Fig. 3).

The average precision of E4 is 0.60. Δ_Q is not more extreme than -27 % in any case and usually less severe than 20 % (Fig. 4). A slight degradation of Δ_Q with increasing ~~temporal resolution~~ frequency of the data can be observed.

4.1.6 Model E4a

E4a results in ~~I_F~~ I_E that are very close to the observed flashiness in all cases: the difference is never more extreme than 0.05 (Fig. 3). $\widehat{I}_{\mathrm{F,det}}$ is often smaller than $I_{\mathrm{F,obs}}$ in the Murg catchment, which, similar as 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 smaller (better) than 0.2 except for the validation period with hourly resolution in both catchments. Like with E4, we can see a tendency for over-fitting with E4a in the Maimai catchment: in the calibration period, reliabilities of 0.02, 0.05 and 0.08 are reached, while the validation results in values of 0.16, 0.16 and 0.23 for daily, 6-hourly and hourly resolutions, respectively (Table B2). A look at the precision metric (Fig. 3) shows that E4a gives 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. 4). Opposed to that, $\widehat{E}_{\mathrm{N,det}}$ is larger 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.

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~~ I_E) compared to E2, which assumes constant correlation (Fig. 3 and 4). For example, the reliability for hourly resolution in the Murg catchment is 0.06 and 0.61 for E3 and E2, respectively (Fig. 3). In contrast to E2, the performance of E3 does not show systematically worse performance for ~~finer temporal resolution of the~~ 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~~ I_E. Figure 6 shows typical results of E2 and E3 w.r.t. streamflow bias, visible as a bias in η, and posterior correlation between heteroscedasticity and correlation parameters a and $\tau_{\max}$. Note also the smaller standard deviation (parameter a) resulting from E3. Additional results about the standardized innovations of η are available in the supplementary material.

Figure 5 compares the predicted hydrographs of E1, E2 and E3a. In this case, allowing for different characteristic correlation times during precipitation events and dry periods (E3a) prevents the problematic behaviour encountered when making the constant correlation assumption. Note that E3a results in better estimates of ~~I_F~~ I_E than E3, since it considers correlation during precipitation events ($\tau_{\min} > 0$). In the Murg Catchment, inferring $\tau_{\min}$ resulted in a degenerative performance for ~~high temporal resolutions~~high-frequency data, which were also linked to higher values of $\tau_{\min}$ (Fig. 7). The posterior estimates of $\tau_{\max}$ depend on the resolution in both catchments. While large $\tau_{\min}$ coincide with the worst reliabilities, large $\tau_{\max}$ were also obtained together with good reliabilities (Fig. 7). The effect of $\tau_{\min}$ 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 $\tau_{\min} > 10\,\mathrm{h}$ and at the same time

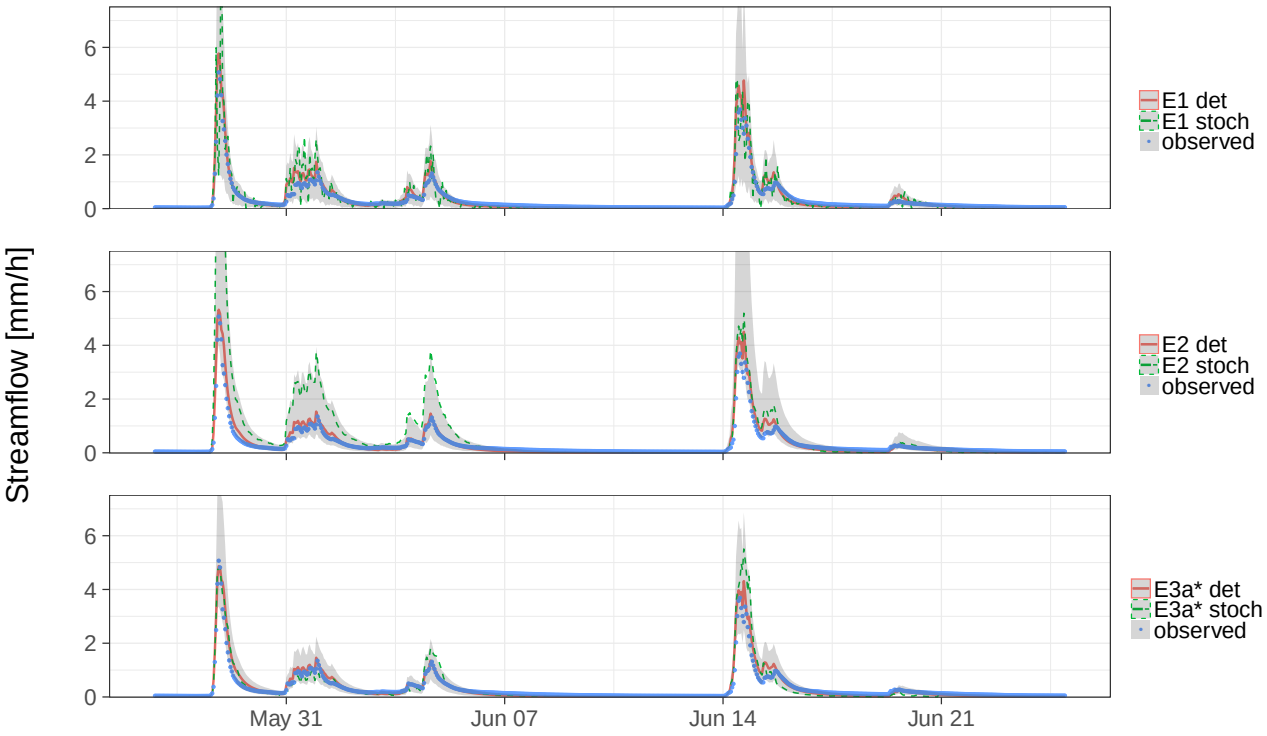

Figure 5. Streamflow predictions ~~for~~ with hourly resolution in the Maimai creek in a part of the validation period (1993). Deterministic predictions with the ~~maximum posterior~~ parameter values at the maximum posterior density are shown together with the 90%-confidence bands and one single stochastic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

$\widehat{E}_{\mathrm{N,det}}$ decreases (not shown), approaching the one of E2.

4.3 Relaxing the assumption of normality

Relaxing the assumption of ~~noramality~~ normality by inferring γ and df (E4 and E4a) had a mixed effect on the numeric performance indices analyzed in this study. When $\tau_{\min} = 0$, including skewness and kurtosis (E4) often led to a better reliability in the calibration period, but a worse reliability in the validation period compared to the assumption of a normal distribution with E3 (Fig. 3). Predictions with E4 were generally less precise than the ones with E3, e.g. Ξ_{reli} was around 0.5 with E3 and 1.0 with E4 for hourly resolution in the Maimai catchment (Fig. 3). When $\tau_{\min}$ 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

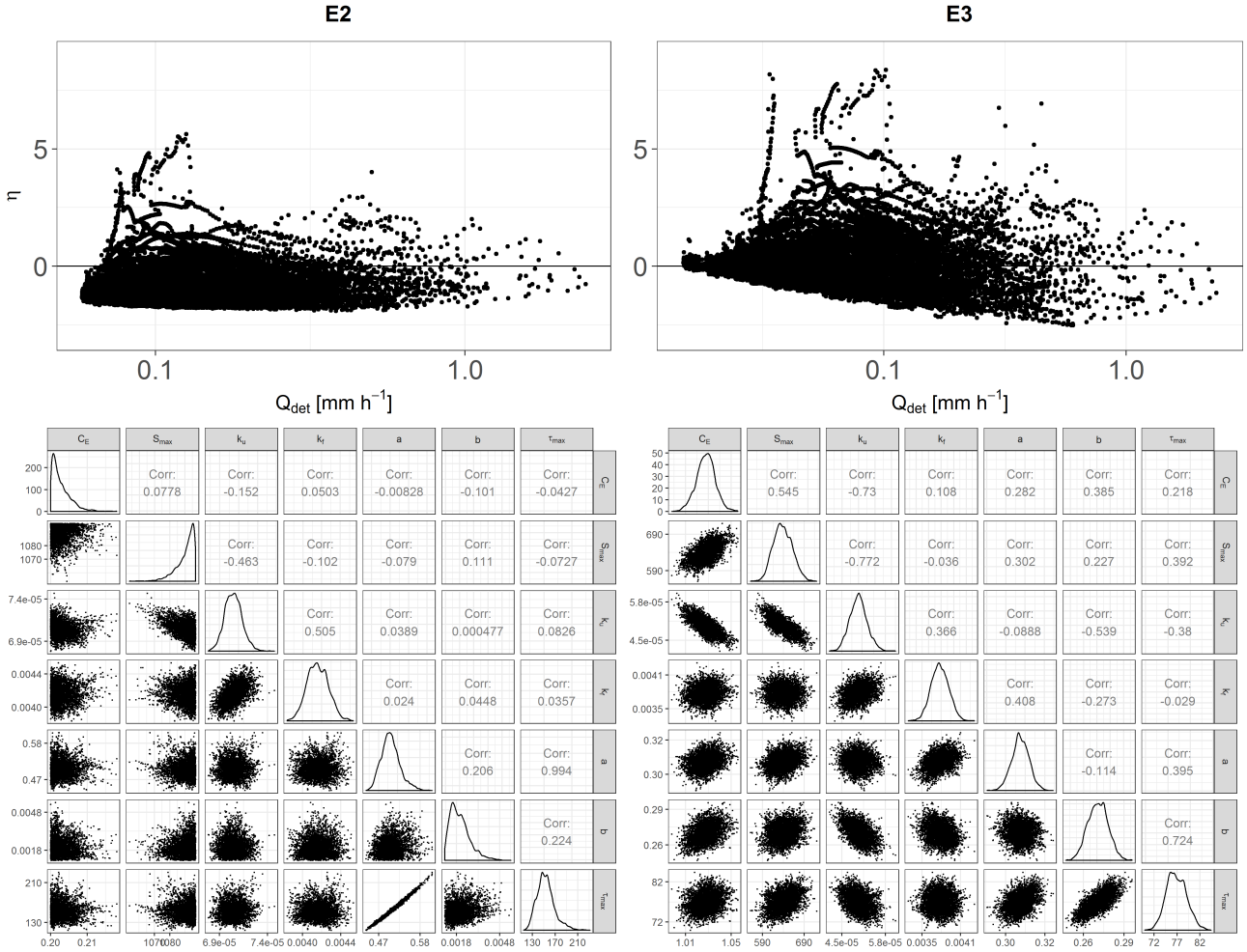

Figure 6. Transformed residuals, η, as a function of modelled streamflow (top) and correlation structure of the posterior parameter sample (bottom) resulting with Error Models E2 (left) and E3 (right) for data with hourly resolution in the Murg catchment.

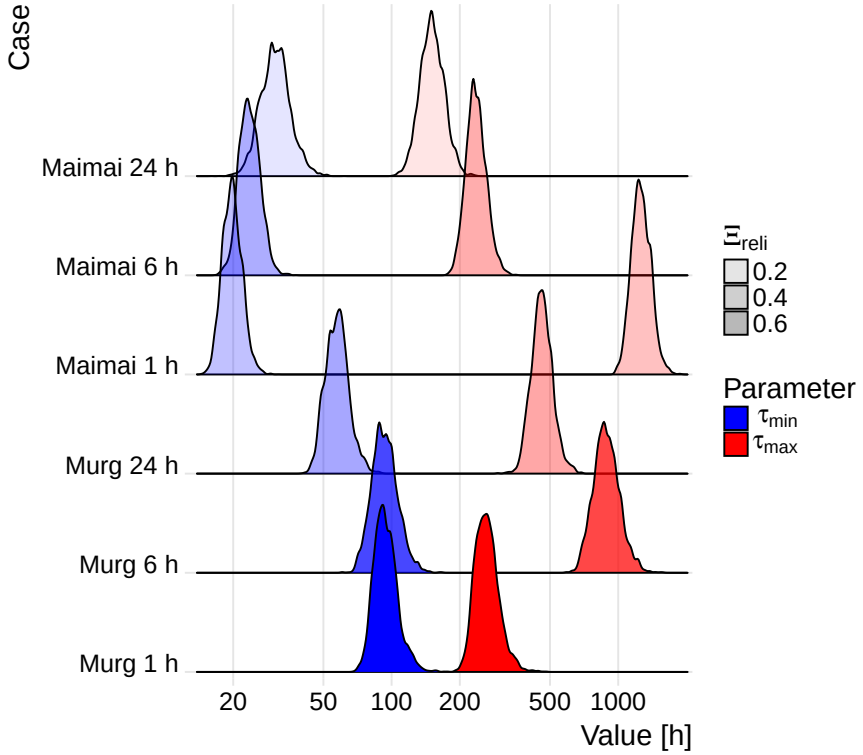

Figure 7. Marginal posterior densities of $\tau_{\min}$ and $\tau_{\max}$, and corresponding reliability measures Ξ_{reli} in the validation period resulting from Error Model E3a in all combinations of catchments and temporal resolutions.

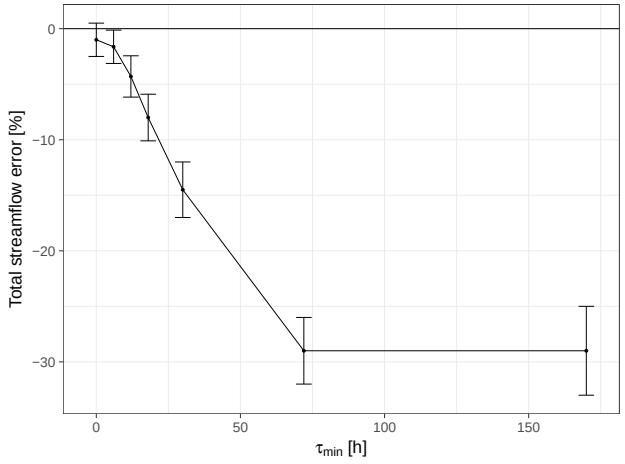

Figure 8. Relationship between the fixed correlation time during precipitation events, $\tau_{\min}$, and the total streamflow error, Δ_Q, for 6-hourly data resolution in the Murg catchment. Each point corresponds to a full inference and prediction procedure. The error bars span two standard deviations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

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 $\widehat{E}_{\mathrm{N,det}}$. For example, E4a resulted in $\gamma \approx 2.5$, $\widehat{E}_{\mathrm{N,det}} = 0.5$ and an underestimation of ~~preak~~ peak flows by the hydrological model for hourly data in the Murg catchment. Inferred df were always at or close to the lower limit of 3, which is indicative of heavy outliers.

Regarding the location of D_Q w.r.t. 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 ~~temporal resolutions~~frequency of the data.

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 $\widehat{Q}_{\mathrm{det}}$ and Q_{obs} in terms of the Nash-Sutcliffe coefficient 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. 9), which translates into unrealistic ~~oscillations~~ fluctuations of the stochastic streamflow predictions (E1 in Fig. 5). 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 analyzed 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. Those are: strong interactions between estimates of hydrological water balance parameters and heteroscedasticity and autocorrelation parameters of the error model (E2 in Fig. 6), smaller ~~E_N~~E_{N}, $\widehat{E}_{\mathrm{N,det}}$, and worse Δ_Q compared to E1. Strong posterior correlations between τ and a coincided with systematic overprediction of streamflow. Evin et al. (2013), who tested an error model similar to E2 on daily data, obtained very similar results in terms of interactions between water balance parameters, heteroscedasticity and ~~correalation~~ correlation 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 ~~finer temporal resolution of the data~~ higher data frequency within two relatively wet catchments.

5.2 (Non-)Stationarity of autocorrelation

Figure 9 visualizes 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 than for the hourly 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 (Fig. 9). Yang et al. (2007) made similar observations. This raises the hypothesis that the neglection of 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 ~~is smaller during periods of rapid change, e.g.~~ of a catchment during precipitation events can be expected to be different from the one during dry weather, the correlation of the errors in streamflow ~~is~~ can be expected to be ~~smaller as well during those times~~different as well. The degree of ~~the reduction 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 others. 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 neglection 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 inference behavior. 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 $\tau_{\max}$ ($\approx 145\,\mathrm{h}$) was much smaller than the original $\tau_{\max}$ ($\approx 1400\,\mathrm{h}$), confirming the hypothesis of reduced correlation time of errors in streamflow during precipitation events.

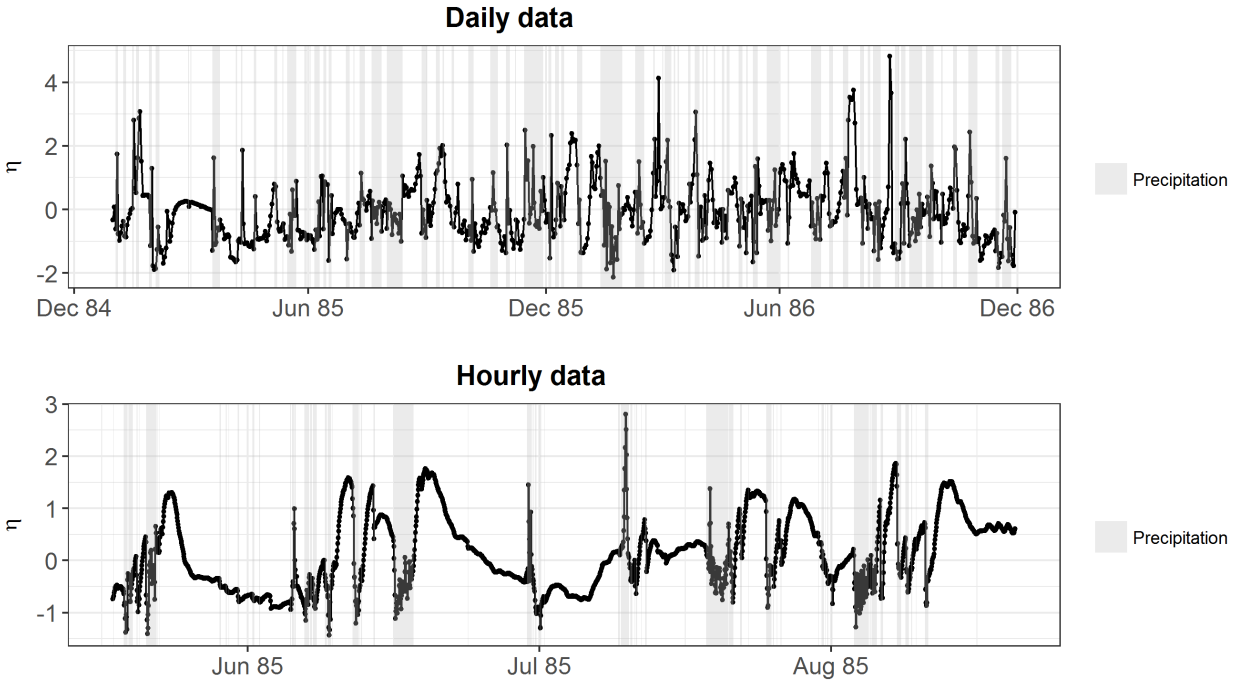

Figure 9. Time series of η corresponding to the parameter values at the maximum posterior ~~parameter set~~ density obtained with E1 in the Maimai catchment for daily ~~resolution (left),~~ and hourly resolution~~(right)~~. Intervals where $P > 0$ are shaded in grey.

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 df, 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 standardized 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 $df = 5$ (E1$^+$). This led to a well-conditioned η and performance metrics that were comparable to or better than the ones 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 of 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 also the distributional parameters of D_Q (e.g. γ or df) could be non-stationary (Wani et al., In preparation).

It is still unclear what the optimal parametrization 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. E.g., there could be a significant lag time 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 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 τ_{min}

The fact that τ_{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 those problems are more related to τ_{min} than to τ_{max}, since higher values of τ_{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 τ_{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 ~~succesful~~ 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~~ 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 df, 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, that 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. 6), 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 analyzed catchments. The previous 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 $Q(t) \sim D_Q(Q_{\mathrm{det}}(t), \psi)$. 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, $E[Q(t)] = Q_{\mathrm{det}}(t)$, mass conservation would be guaranteed (since the applied hydrological model conserves mass). In our experience, however, such distributions lead to unsatisfactory fits (results not shown). An extreme right-skew is needed to account for cases where $Q_{\mathrm{obs}}(t) \gg E[Q(t)]$, i.e. when observed streamflow is several orders of magnitude larger than modelled streamflow, which can happen if the latter approaches zero. Such an extreme right-skew is unrealistic in case of larger modelled streamflows, where it is very unlikely that the observations are several orders of magnitude larger than a well-calibrated hydrological model. Thus, the non-negativity of discharge observations (for non-tidal rivers) makes it in our experience practically nearly impossible to keep mass balances at very low discharge if there is a considerable observation error. A possible solution to this problem might be non-stationary skewness or kurtosis of D_Q (Wani et al., In preparation), e.g. imposing an inverse relation between skewness or kurtosis and Q_{det}.

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 characterizing 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 ~~likelihood~~ 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 invalidate 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 Nash-Sutcliffe Efficiency ~~goes~~ 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 ~~goes~~ decreases from 0.83 to 0.79).

3. Since rapid changes in a catchments storage ~~reduce~~ change its memory, errors in streamflow are expected to ~~be less correlated~~ show different correlations during precipitation events ~~than during~~ 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 ~~likeli~~ 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. ~~Accounting~~ 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 can lead to better fulfilled distributional assumptions about the errors~~in case of low temporal resolution of the data. For higher resolutions, however~~. 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 w.r.t. the shape of the distribution can actually lead to less accurate representation of the observed distribution.

These results contribute to a better characterization of the residual errors of deterministic hydrological models. However, some questions remain. For example, it is still unclear how the non-stationary autocorrelation should ideally be parametrized. 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, it could

be investigated whether distributions other than the Gaussian and the skewed Student's t are more ~~appropirate~~appropriate or whether a different model for the temporal correlation would perform better. Overall, this study confirms previously encountered difficulties in finding a parametrization of an additive error term that adequately describes the effects of intrinsic stochasticity.

Appendix A: Derivation of the likelihood function

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

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

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

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

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

This corresponds to the first sub-equation of Eq. (7). The order of the factors was changed in Eq. (7) to emphasize the product of the marginal distribution f_{D_Q} with a modification ~~facor~~ factor that tends to unity if ~~$t_{i+1} - t_i$~~ $t_i - t_{i-1}$ 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

Appendix C: ~~Synthetic case study: inferring known true parameters~~

~~To check if the implemented likelihood and sampling functions are inverses of each other, we produce a streamflow sample with known parameters according to the procedure outlined in Sect. 2.3. Table **??** shows the results when trying to re-infer those known parameters with the error models presented in this study. In all cases, the true value of the parameters are inside~~

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_{prec} indicate better results. * : smoothing $P_{\text{err}}(t)$ with a moving average window of size 5 h before applying Eq. 11.$^{\sim}$denotes the option where $\text{mode}(D_Q) = Q_{\text{det}}$. $^{+}$ means that $\gamma = 1.5$ and $df = 5$ was fixed.

	Calibration									Validation							
Case	Ξ_{reli}	Ω_{prec}	$\widehat{E}_{\text{N,det}}$	~~E_N~~ E_{N}	~~$\widehat{\Delta}_{Q,\text{det}}$~~ $\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	~~I_F~~ I_{F}	$I_{\text{F,obs}}$	Ξ_{reli}	Ω_{prec}	$\widehat{E}_{\text{N,det}}$	~~E_N~~ E_{N}	~~$\widehat{\Delta}_{Q,\text{det}}$~~ $\widehat{\Delta}_{Q,\text{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\text{F,det}}$	~~I_F~~ I_{F}
24h E1	0.17	0.32	0.83	0.68(0.04)	-10	-10(2.3)	0.29	0.36(0.01)	0.31	0.2	0.31	0.81	0.63(0.04)	-8	-8(2.4)	0.29	0.36(0.01)
24h E2	0.3	0.43	0.76	0.48(0.1)	-23	-23(5)	0.29	0.34(0.01)	0.31	0.32	0.42	0.75	0.41(0.1)	-21	-22(5.3)	0.28	0.33(0.01)
24h E3	0.09	0.35	0.84	0.65(0.04)	-3	-3(1.6)	0.3	0.45(0.01)	0.31	0.09	0.34	0.8	0.59(0.04)	-1	-1(1.7)	0.29	0.44(0.02)
24h E3a	0.21	0.37	0.83	0.62(0.06)	-11	-12(3.5)	0.29	0.36(0.01)	0.31	0.21	0.37	0.8	0.56(0.05)	-8	-10(3.7)	0.29	0.35(0.01)
24h $\widetilde{\text{E4}}$	0.04	0.35	0.83	0.66(0.1)	5	-1(1.7)	0.27	0.41(0.02)	0.31	0.09	0.34	0.78	0.59(0.16)	6	1(1.9)	0.26	0.4(0.02)
24h E4	0.03	0.38	0.84	0.65(0.19)	0	0(1.6)	0.24	0.39(0.02)	0.31	0.07	0.37	0.78	0.57(0.1)	2	1(1.8)	0.23	0.38(0.02)
24h $\widetilde{\text{E4}}$a	0.04	0.38	0.76	0.6(0.16)	16	-3(3.7)	0.22	0.29(0.01)	0.31	0.15	0.38	0.64	0.48(0.25)	18	-2(4)	0.2	0.28(0.01)
24h E4a	0.03	0.35	0.81	0.66(0.37)	5	4(2.6)	0.2	0.28(0.02)	0.31	0.06	0.35	0.72	0.56(1.4)	6	5(2.8)	0.19	0.27(0.02)
6h E1	0.13	0.38	0.81	0.59(0.03)	-4	-4(0.8)	0.12	0.44(0.01)	0.16	0.12	0.37	0.8	0.57(0.02)	-2	-2(0.8)	0.12	0.43(0.01)
6h E2	0.4	0.56	0.6	0.13(0.15)	-34	-40(5.7)	0.14	0.18(0)	0.16	0.35	0.54	0.6	0.05(0.14)	-30	-35(5.7)	0.14	0.18(0)
6h E3	0.1	0.42	0.76	0.5(0.04)	-3	-3(1.4)	0.15	0.36(0.01)	0.16	0.04	0.41	0.76	0.48(0.03)	0	0(1.5)	0.14	0.36(0.01)
6h E3a	0.52	0.63	0.55	-0.03(0.17)	-41	-50(7.3)	0.14	0.18(0)	0.16	0.46	0.61	0.52	-0.18(0.2)	-36	-45(7.5)	0.14	0.18(0)
6h $\widetilde{\text{E4}}$	0.05	0.38	0.79	0.62(0.1)	7	-2(1.4)	0.1	0.27(0.01)	0.16	0.1	0.37	0.74	0.56(0.08)	8	0(1.4)	0.1	0.27(0.01)
6h E4	0.07	0.43	0.79	0.59(0.1)	-2	-2(1.4)	0.08	0.27(0.01)	0.16	0.09	0.41	0.75	0.52(0.13)	0	0(1.5)	0.08	0.27(0.01)
6h $\widetilde{\text{E4}}$a	0.05	0.4	0.63	0.51(0.64)	25	2(3.1)	0.07	0.14(0.01)	0.16	0.08	0.38	0.45	0.35(0.19)	27	5(3.2)	0.06	0.13(0.01)
6h E4a	0.11	0.31	0.67	0.58(0.07)	12	12(2.1)	0.06	0.13(0)	0.16	0.07	0.3	0.55	0.45(0.08)	13	13(2.2)	0.05	0.12(0)
1h E1	0.1	0.38	0.79	0.54(0.01)	-1	-1(0.4)	0.03	0.43(0)	0.05	0.13	0.39	0.78	0.58(0.02)	-2	-2(0.5)	0.03	0.43(0.01)
1h E1^{+}	0.06	0.33	0.75	0.55(0.02)	14	-4(0.4)	0.02	0.33(0)	0.05	0.08	0.33	0.74	0.58(0.02)	12	-6(0.5)	0.02	0.33(0)
1h E2	0.61	0.73	0.09	-0.9(0.33)	-61	-76(7.8)	0.04	0.06(0)	0.05	0.61	0.7	0.34	-0.28(0.24)	-56	-70(8.8)	0.04	0.06(0)
1h E2^{+}	0.48	0.2	0.5	0.43(0.05)	30	20(1.8)	0.01	0.05(0)	0.05	0.24	0.22	0.59	0.53(0.05)	22	10(2.3)	0.01	0.05(0)
1h E3	0.07	0.42	0.69	0.45(0.02)	5	-1(1.3)	0.04	0.22(0)	0.05	0.06	0.4	0.69	0.52(0.02)	7	2(1.6)	0.04	0.24(0.01)
1h E3a	0.6	0.74	0.15	-0.7(0.29)	-61	-73(8.8)	0.03	0.06(0)	0.05	0.64	0.75	0.36	-0.31(0.25)	-56	-74(10.7)	0.03	0.06(0)
1h E3a*	0.62	0.74	0.21	-0.6(0.26)	-62	-72(8.8)	0.03	0.06(0)	0.05	0.65	0.75	0.39	-0.27(0.22)	-56	-75(9.4)	0.03	0.06(0)
1h $\widetilde{\text{E4}}$	0.49	0.96	0.45	-0.29(0.26)	30	-47(4.3)	0.01	0.18(0)	0.05	0.51	0.93	0.48	-0.02(0.23)	28	-45(4.9)	0.01	0.19(0)
1h E4	0.2	0.59	0.65	0.41(0.26)	-8	-9(2.5)	0.01	0.15(0)	0.05	0.24	0.56	0.63	0.47(0.07)	-8	-9(2.5)	0.01	0.16(0)
1h $\widetilde{\text{E4}}$a	0.85	1.85	0.49	-5.58(0.97)	15	-205(15.1)	0.01	0.07(0)	0.05	0.87	1.78	0.49	-3.8(0.83)	13	-200(17.1)	0.01	0.08(0)
1h E4a	0.18	0.38	0.5	0.4(0.03)	4	4(2.1)	0.01	0.06(0)	0.05	0.22	0.36	0.49	0.41(0.03)	4	3(2.2)	0.01	0.06(0)

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_{prec} indicate better results. * : smoothing $P_{\mathrm{err}}(t)$ with a moving average window of size 5 h before applying Eq. 11.~denotes the option where $\mathrm{mode}(D_Q) = Q_{\mathrm{det}}$. †: $P_{\mathrm{err}} \neq P$.

Case	Calibration									Validation							
	Ξ_{reli}	Ω_{prec}	$\widehat{E}_{\mathrm{N,det}}$	~~E_N~~ E_{N}	~~$\widehat{\Delta}_{\mathrm{Q,det}}$~~ $\widehat{\Delta}_{Q,\mathrm{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\mathrm{F,det}}$	~~I_F~~ I_{F}	$I_{\mathrm{F,obs}}$	Ξ_{reli}	Ω_{prec}	$\widehat{E}_{\mathrm{N,det}}$	~~E_N~~ E_{N}	~~$\widehat{\Delta}_{\mathrm{Q,det}}$~~ $\widehat{\Delta}_{Q,\mathrm{det}}$ [%]	Δ_Q [%]	$\widehat{I}_{\mathrm{F,det}}$	~~I_F~~ I_{F}
24h E1	0.09	0.42	0.92	0.73(0.06)	-8	-8(3.7)	0.77	0.88(0.03)	0.83	0.09	0.4	0.91	0.7(0.08)	-3	-4(3.9)	0.84	0.94(0.04)
24h E2	0.25	0.52	0.91	0.62(0.1)	-17	-20(7.4)	0.74	0.8(0.03)	0.83	0.07	0.49	0.9	0.59(0.13)	-11	-13(7.9)	0.81	0.85(0.03)
24h E3	0.11	0.45	0.91	0.7(0.08)	-11	-11(4.1)	0.79	0.89(0.04)	0.83	0.1	0.43	0.9	0.65(0.1)	-6	-5(4.5)	0.87	0.95(0.04)
24h E3a	0.22	0.49	0.91	0.64(0.09)	-16	-18(6.4)	0.75	0.82(0.03)	0.83	0.04	0.47	0.9	0.62(0.11)	-10	-11(6.3)	0.82	0.88(0.04)
24h $\widetilde{\mathrm{E4}}$	0.05	0.56	0.92	0.53(0.24)	-6	-16(6.2)	0.81	0.93(0.05)	0.83	0.14	0.54	0.91	0.48(0.29)	-1	-12(6.8)	0.88	0.99(0.05)
24h E4	0.02	0.57	0.92	0.6(0.27)	-9	-9(5.5)	0.78	0.9(0.04)	0.83	0.19	0.54	0.91	0.57(0.33)	-4	-3(5.7)	0.85	0.95(0.05)
24h $\widetilde{\mathrm{E4}}$a	0.07	0.62	0.92	0.44(0.37)	-5	-24(8.4)	0.8	0.88(0.04)	0.83	0.07	0.63	0.91	0.37(2.74)	-1	-19(10.9)	0.88	0.95(0.05)
24h E4a	0.02	0.56	0.92	0.6(0.25)	-10	-10(6.6)	0.77	0.85(0.04)	0.83	0.16	0.56	0.91	0.55(0.94)	-5	-6(7.5)	0.85	0.91(0.05)
6h E1	0.09	0.45	0.89	0.69(0.05)	-11	-11(2.2)	0.4	0.63(0.02)	0.46	0.12	0.42	0.88	0.68(0.05)	-5	-5(2.5)	0.45	0.65(0.03)
6h E2	0.47	0.67	0.83	0.34(0.19)	-37	-51(9.3)	0.36	0.39(0.01)	0.46	0.27	0.62	0.83	0.37(0.2)	-27	-40(11.2)	0.39	0.43(0.01)
6h E3	0.15	0.49	0.84	0.58(0.07)	-17	-18(2.8)	0.45	0.61(0.02)	0.46	0.1	0.46	0.82	0.54(0.09)	-10	-10(3.3)	0.5	0.65(0.03)
6h E3a	0.11	0.56	0.85	0.5(0.12)	-14	-25(6)	0.38	0.45(0.02)	0.46	0.19	0.51	0.84	0.49(0.14)	-6	-16(7)	0.42	0.49(0.02)
6h $\widetilde{\mathrm{E4}}$	0.08	0.59	0.89	0.52(0.14)	-13	-14(3.6)	0.42	0.65(0.03)	0.46	0.2	0.56	0.87	0.48(0.21)	-7	-7(4.2)	0.46	0.67(0.03)
6h E4	0.03	0.72	0.89	0.39(0.27)	-15	-16(4.1)	0.4	0.65(0.03)	0.46	0.18	0.67	0.87	0.39(0.88)	-9	-9(4.7)	0.44	0.67(0.04)
6h $\widetilde{\mathrm{E4}}$a	0.06	0.92	0.89	-0.19(1.76)	-1	-46(12.5)	0.38	0.47(0.02)	0.46	0.14	0.84	0.88	-0.08(1.66)	5	-37(12.2)	0.42	0.51(0.03)
6h E4a	0.05	1.45	0.83	-0.45(3.16)	-29	-39(19.3)	0.33	0.44(0.02)	0.46	0.16	1.27	0.83	-0.34(5.2)	-19	-26(20.6)	0.36	0.46(0.03)
1h E1	0.09	0.56	0.84	0.48(0.04)	-19	-19(1.2)	0.14	0.56(0.01)	0.13	0.15	0.52	0.8	0.41(0.06)	-11	-12(1.4)	0.15	0.56(0.01)
1h E2	0.29	0.87	0.78	-1.19(0.73)	-26	-96(17.2)	0.12	0.13(0)	0.13	0.14	0.77	0.71	-1.38(0.92)	-15	-77(17.1)	0.13	0.14(0)
1h E3	0.09	0.58	0.78	0.43(0.04)	-21	-26(1.9)	0.13	0.38(0.01)	0.13	0.29	0.52	0.71	0.33(0.07)	-11	-14(2.3)	0.14	0.39(0.01)
1h E3*	0.16	0.5	0.83	0.62(0.02)	-16	-22(2)	0.11	0.32(0.01)	0.13	0.18	0.44	0.8	0.59(0.04)	-6	-10(2.2)	0.12	0.32(0.01)
1h E3*†	0.14	0.72	0.78	0.02(0.24)	-30	-51(6.7)	0.11	0.26(0.01)	0.13	0.14	0.65	0.73	-0.14(0.29)	-19	-41(7.3)	0.13	0.24(0.01)
1h E3a*	0.24	0.5	0.81	0.51(0.1)	-20	-36(5.8)	0.11	0.13(0)	0.13	0.14	0.44	0.77	0.43(0.15)	-10	-24(6.2)	0.13	0.15(0)
1h $\widetilde{\mathrm{E4}}$	0.12	0.82	0.83	-0.11(0.46)	-16	-34(3)	0.13	0.46(0.02)	0.13	0.3	0.77	0.8	-0.2(1.21)	-9	-25(3.7)	0.14	0.47(0.02)
1h $\widetilde{\mathrm{E4}}^\star$	0.07	0.63	0.86	0.31(0.43)	-8	-24(2.2)	0.12	0.42(0.02)	0.13	0.27	0.59	0.86	0.3(0.56)	-1	-16(2.6)	0.13	0.43(0.02)
1h E4*	0.11	1.02	0.8	-0.17(0.61)	-27	-27(3.9)	0.1	0.5(0.02)	0.13	0.33	0.94	0.78	-0.22(0.93)	-16	-16(4.8)	0.11	0.51(0.03)
1h $\widetilde{\mathrm{E4}}$a*	0.2	0.4	0.72	0.58(0.33)	18	1(4.3)	0.08	0.12(0)	0.13	0.32	0.4	0.71	0.58(0.27)	19	2(4.8)	0.08	0.12(0.01)
1h E4a*	0.08	1.2	0.78	-0.27(1.76)	-35	-42(13.9)	0.1	0.16(0.01)	0.13	0.23	1.08	0.76	-0.32(1.8)	-22	-28(15.7)	0.11	0.17(0.01)

the posterior 95%-confidence intervals. Synthetic case study based on parameter values obtained for the Murg river with daily resolution. ML = maximum likelihood estimation, CI = confidence interval. Param.Unit True valueML95%-CIML95%-CIML95%-CIML95 - 1.031.0431.0271.020.981.02 1.01.041.0071.03 1.004 1.041.01.0631.071.041.0751.0581.06$S_{\max}$ mm 34535232933629832 5 302 3303003303073383023703733463653563 68k_u h^{-1} 8.5e-58.59e-57.7e-5 9.76e-57.44e-59.5e-5 8.0e-5 7.95e-56.09e-59.87e-58.14e-58 h^{-1}7.59e-48.31e-47.06e-4 8.82e-57.09e-58.29e-46.6e-4 6.77e-45.39e-48.83e-46.81e-47.72e-46.65e-49.88e-41.19e-49.8e-4 7.96e-41.00e- - 2.042.03 1.911.82 1.621.93 1.81 2.121.921.961.791.951.772.15 2.162.072.272.142.11b - 2.0e-22.0e-21.2e-2 0.0170.0111.66e-24.6e-3 2.45e-21.43e-23.44e-21.77e-21.63e-21.05e-23.4e-20.0423.0e-24.23e-25.17e-22.99e-2$\tau_{\max}$ h 393- - 297 255385 335 41434737432635730 h 48- - - - -47.241.3--44.638.6- -57.7-52.7γ - 1.2- - - - - - -1.211.151.161.10- - - -1.281.23df - 7.39- - - - - - -6.093.979.485.63- - - -10.218.3

Appendix C: Specific error models

C1 Normal distribution

$$D_Q = \mathrm{N}(\mu, \sigma)$$
$$\mu(Q_{\mathrm{det}}) = Q_{\mathrm{det}} \quad , \quad \sigma(Q_{\mathrm{det}}, a, b, c) = aQ_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + bQ_0 \quad , \quad \psi = (a, b, c) \tag{C1}$$

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~~t distribution~~'s *t*-distribution

$$D_Q = T_{df,\sigma}(\mu, \sigma, df)$$
$$\mu(Q_{\mathrm{det}}) = Q_{\mathrm{det}} \quad , \quad \sigma_{T_{df}} = aQ_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + bQ_0 \quad , \quad \psi = (a, b, c) \tag{C2}$$

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

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

and

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

Note that the degrees of freedom, df, 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~~t distribution~~'s *t*-distribution

$$
\begin{aligned}
D_Q &= \mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}](Q_{\mathrm{det}}, \sigma, df, \gamma) \\
\sigma_{\mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}]} &= aQ_0 \left(\frac{Q_{\mathrm{det}}}{Q_0}\right)^c + bQ_0 \quad , \quad \boldsymbol{\psi} = (a, b, c)
\end{aligned}
\tag{C5}
$$

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

$$
f_{\mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}]}(x) = \begin{cases}
\dfrac{2}{\gamma + \frac{1}{\gamma}} f_{\mathrm{T}_{df,\sigma}}(\gamma x) = \dfrac{2}{\gamma + \frac{1}{\gamma}} \dfrac{1}{\sigma} \sqrt{\dfrac{df}{df-2}}\, f_{\mathrm{T}_{df}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{df}{df-2}}\,\gamma x\right) & \text{for } x \leq 0 \\
\dfrac{2}{\gamma + \frac{1}{\gamma}} f_{\mathrm{T}_{df,\sigma}}\left(\dfrac{x}{\gamma}\right) = \dfrac{2}{\gamma + \frac{1}{\gamma}} \dfrac{1}{\sigma} \sqrt{\dfrac{df}{df-2}}\, f_{\mathrm{T}_{df}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{df}{df-2}}\,\dfrac{x}{\gamma}\right) & \text{for } x \geq 0\,.
\end{cases}
\tag{C6}
$$

and

$$
F_{\mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}]}(x) = \begin{cases}
\dfrac{2}{1+\gamma^2} F_{\mathrm{T}_{df,\sigma}}(\gamma x) = \dfrac{2}{1+\gamma^2} F_{\mathrm{T}_{df}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{df}{df-2}}\,\gamma x\right) & \text{for } x \leq 0 \\
\dfrac{1}{1+\gamma^2} + \dfrac{2}{1+\frac{1}{\gamma^2}}\left(F_{\mathrm{T}_{df,\sigma}}\left(\dfrac{x}{\gamma}\right) - \dfrac{1}{2}\right) & \\
\quad = \dfrac{1}{1+\gamma^2} + \dfrac{2}{1+\frac{1}{\gamma^2}}\left(F_{\mathrm{T}_{df}}\left(\dfrac{1}{\sigma}\sqrt{\dfrac{df}{df-2}}\,\dfrac{x}{\gamma}\right) - \dfrac{1}{2}\right) & \text{for } x \geq 0\,.
\end{cases}
\tag{C7}
$$

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

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

and:

$$\sigma^2_{\mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}]} = \frac{\gamma^3 + \frac{1}{\gamma^3}}{\gamma + \frac{1}{\gamma}}\,\sigma^2 - \mu^2_{\mathrm{sk}_\gamma[\mathrm{T}_{df,\sigma}]}$$

$$= \left(\frac{\gamma^3 + \frac{1}{\gamma^3}}{\gamma + \frac{1}{\gamma}} - 4 \left(\frac{\gamma^2 - \frac{1}{\gamma^2}}{\gamma + \frac{1}{\gamma}} \right)^2 \frac{df\,(df-2)}{(df-1)^2} \frac{\Gamma^2\left(\frac{df+1}{2}\right)}{\pi\ df\ \Gamma^2\left(\frac{df}{2}\right)} \right) \sigma^2 \quad . \tag{C9}$$

To shift the distribution we can evaluate

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

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

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

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 ~~maximum posterior probability~~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}.
ψ	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.
$\tau_{\min}$	Minimum value of τ in the cases where τ is a function of P_{err} and therefore of time.
$\tau_{\max}$	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.
$\mathrm{T}(df, \sigma)$	Rescaled ~~Student-t distribution~~ Student's *t*-distribution with df degrees of freedom and standard deviation σ.
$\mathrm{SKT}(\mu, \sigma, df)$	Shifted and rescaled skewed ~~Student-t distribution~~ Student's *t*-distribution with mean μ, standard deviation σ, and df degrees of freedom.
F_I	The median of the Flashiness Indices (Baker et al., 2004) of all the individual model realisations constituting a sample of model outputs.
$\widehat{F}_{\mathrm{I,det}}$	The Flashiness Index (Baker et al., 2004) of $\widehat{Q}_{\text{det}}$.
$F_{\mathrm{I,obs}}$	The Flashiness Index (Baker et al., 2004) of Q_{obs}.
E_{N}	The median of the Nash-Sutcliffe Indices (Nash and Sutcliffe, 1970) of all the individual model realisations constituting a sample of model outputs.
$\widehat{E}_{\mathrm{N,det}}$	The Nash-Sutcliffe Index (Nash and Sutcliffe, 1970) of $\widehat{Q}_{\text{det}}$.
Δ_Q	The median of the relative errors in cumulative streamflow of all the individual model realisations constituting a sample of model outputs.
$\widehat{\Delta}_{Q,\text{det}}$	The relative error in cumulative streamflow of $\widehat{Q}_{\text{det}}$.
Ξ_{reli}	Reliability metric (McInerney et al., 2017)
Ω_{prec}	Precision metric (McInerney et al., 2017)
OU-process	Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930).

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 (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 an anonymous reviewer for their feedback and their help in improving this manuscript.

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, Journal of the American Water Resources Association, 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, Hydrology and Earth System Sciences, 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 Resources Research, 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, Advances in Water Resources, 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, Hydrological Processes, 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 '07), IEEE, 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, Journal of Hydrology, 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, Hydrology and Earth System Sciences, 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 Resources Research, 24, 1163–1173, https://doi.org/10.1029/wr024i007p01163, 1988.

Evin, G., Kavetski, D., Thyer, M., and G., K.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resources Research, 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 Resources Research, 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 Resources Research, 47, 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 Resources Research, https://doi.org/10.1002/2017wr021616, 2018.

Fernandez, C. and Steel, M. F. J.: On Bayesian Modeling of Fat Tails and Skewness, Journal of the American Statistical Association, 93, 359–371, 1998.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J.: emcee: The MCMC hammer, Publications of the Astronomical Society of the Pacific, 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 Resources Research, 32, 2161–2173, https://doi.org/10.1029/95wr03723, 1996.

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resources Research, 47, 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, pp. 49–68, American Geophysical Union, 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 Resources Research, 42, 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, Journal of Hydrology, 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 Resources Research, 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, 2002.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, The Annals of Mathematical Statistics, 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 Resources Research, 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, Journal of Hydrology, 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 Resources Research, 53, 2199–2239, https://doi.org/10.1002/2016wr019168, 2017.

Meteoschweiz: https://www.meteoschweiz.admin.ch/home/service-und-publikationen/beratung-und-service/datenportal-fuer-experten.html, 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 Resources Research, 40, https://doi.org/10.1029/2003wr002540, 2004.

Montanari, A. and Koutsoyiannis, D.: Modeling and mitigating natural hazards: Stationarity is immortal!, Water Resources Research, 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, Journal of Hydrology, 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 Resources Research, 46, W05 521, doi:10.1029/2009WR008 328, 2010.

Schaefli, B., Talamba, D. B., and Musy, A.: Quantifying hydrological modeling errors through a mixture of normal distributions, Journal of Hydrology, 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, Hydrology and Earth System Sciences Discussions, 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, Hydrological Processes, 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 Resources Research, 46, W10 531, doi:10.1029/2009WR008 933, 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 Resources Research, 38, 1241, doi:10.1029/2001WR000 978, 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 Resources Research, 46, https://doi.org/10.1029/2010wr009514, 2010.

Sun, X., Mein, R., Keenan, T., and Elliott, J.: Flood estimation using radar and raingauge data, Journal of Hydrology, 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 Resources Research, 45, https://doi.org/10.1029/2008wr006825, 2009.

Uhlenbeck, G. E. and Ornstein, L. S.: On the Theory of the Brownian Motion, Physical Review, 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, Environmental Modelling & Software, 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, Hydrology and Earth System Sciences, 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, In preparation.

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