Reply to Referee #1, Jasper Vrugt

AC: Authors comment

Intended changes to the manuscript are highlighted in italic font.

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 will mention this more explicitly in the revised manuscript*.

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 will rephrase 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 will mention this shortcoming of previous approaches more explicitly in the paragraph on page 3, line 20. Accordingly, we will expand 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 will reverse the order of Eq.(2) and Eq.(3) and we will include some more clarifying text about Eq.(3).

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 will include a corresponding sentence in the next version of the manuscript. We will also more clearly discuss 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). Both options may be useful, depending on the circumstances (e.g. for ephemeral catchments the latter option could be preferable, since a non-zero probability of zero discharge may be very important there).*

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 intended for comment 3, we hope that this will become 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 will state 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 will replace 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 will mention this more explicitly in the next version*.

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 will replace eta_trans by eta(ti), i.e. we will substitute Eq. (2) into Eq. (7). Since the dependence of eta(t_i) on Q(t_i) is then not explicit anymore, we will add a statement about that dependence and refer to Eq. (2).*

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" indstead 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 will include a corresponding statement about Eq. (4) in the next 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 will refer explicitly to Eq. (1) and also modify the index "i" in Eq. (7) so that it has the same meaning as in Eq. (8).*

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 will clarify this in the next version and replace 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 will include the specific settings used for sampling with this ensemble sampler in the next version of the manuscript*.
>
> 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 will check and improve regular versus italics fonts in equations throughout the manuscript*.

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".
>
> *In the next version we will change 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 will include more explanations and references to the numerical scheme used for integrating the equations in the next version.*

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 will include the above mentioned explanations as to why those parameters were kept fixed in the next version, but we will 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 will include a reference to that discussion in the next version of the manuscript. We will say which aspects of the approach of Scharnagl et al. (2015) are similar to this study, but we will also highlight the important differences.

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.

In the revised version, we will add a short paragraph 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 will change the order of the sections in the next version of the manuscript.

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. *The next version of the manuscript will be changed accordingly*.

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 will modify 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.

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 will include such a plot in the Appendix in the next version.

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 will include some plots of the priors and the posteriors in the appendix, and we will compute the KL divergence and include that information in the appendix, too.

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 will try to make these plots more easily interpretable by changing some of the symbols or by summarizing some of the dimensions.

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 will enlarge the panels of Figure 5 horizontally*. *Additionally, we will also enlarge 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 will modify the text at several places to clearly state that we are only dealing with hydrological models that predict discharge based on given rainfall.

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 will expand 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.

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.