Reply to Referee #2

RC: Referee comment

AC: Authors comment

Intended changes to the manuscript are highlighted in italic font.

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 will include 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 will include it in the title of the paper, which will be changed to something similar to: "A new framework for likelihood functions of 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 the next version of the manuscript, we will improve 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.

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 will complement 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 will add 2 arrows in Figure 3, clarifying that smaller reliability and precision values mean better results. We will also add 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 will mention in the next version of the manuscript 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 will add 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 will add 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 will add a statement on page 24 line 16 and adapt finding 3 of the conclusions to say that in principle also higher correlation of errors during precipitation events is possible, *but we will 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 will correct them in the next version of the manuscript.