Reviewer #1

Summary and Recommendation

Nearing et al. test how near-real time streamflow observations can effectively be used in Long Short-Term Memory (LSTM) rainfall runoff models. They compare an autoregression (AR) approach with a data assimilation (DA) approach and test, additionally, how sensitive AR is to random gaps in the data. The manuscript (MS) is easy to follow, well-structured and suits the scope of HESS. I particularly liked the comprehensive appendix in combination with a short MS. I think that this MS can be published after some minor revisions and provide only some smaller comments and questions below.

Sincerely

Ralf Loritz

Questions and comments:

- Reading the MS I would have like to see a couple of detailed results from three or four catchments where the AR or DA worked particular good or bad and what the (hydrological and ML) reason for this might be (to underpin the discussion of Appendix F and G). For instance, what could be the reason that DA and AR reduces the predictive performance of a few of your models (Fig. F2)? You state that: "*We are unsure of the reason for this, but it warrants further exploration.*" (Line 352) maybe zooming into one of the catchments could help to give a better explanation.

Thank you for the suggestion. We do agree with the notion of this comment. Indeed, We tried this extensively before submitting the manuscript and did not find a cohesive story to tell. Of course we looked at hydrographs extensively, there just isn't an interesting story to tell there.

- I find it a bit unrealistic how you added the missing data. I would assume that a broken gauging station is not working for a couple of days or maybe weeks in a row and wonder how this would alter your results (e.g. all streamflow data available for training but then two weeks or more only simulated data during testing with a closer focus on particular that period and not the entire testing period).

We did two things. First we added sequences of missing data (capped at length 10 days, since beyond this length, the model generally does not retain information from the past streamflow inputs or assimilation data. Also, we will include an "ungauged basin" analysis,

where we test both methods (DA, AR) on catchments where there is no gauge data at all during training and/or inference.

- Showing how the variance or the Shannon entropy changes of your simulations in addition to the median would be interesting (Fig.1 and Fig.3). If it remains constant, I would mention that the spread of the predictions is not affected by the data availability.

Yes, this is a strange oversight on our part. We included CDF plots in the revised manuscript.

Personal comment: Three of the seven Co-Authors have presumingly not contributed to this “technical note” as they are not mentioned in the author contribution section.

Thank you. This oversight was fixed.

Reviewer #2

The technical note compares two different techniques for using near-real-time streamflow observations to improve operational streamflow forecasts from LSTM rainfall-runoff models. The first technique ("autoregression", AR) adds lagged streamflow observations as predictor in the model. The second technique uses variational data assimilation (DA) to update model states within an assimilation window. The two techniques are compared on the CAMELS dataset, including experiments that artificially remove data to simulate scenarios with missing streamflow data.

The paper is generally well written, concise and to the point. The comparison between AR and DA is an interesting and novel contribution to the literature.

Comments:

1. The main conclusion is that "AR significantly out-performed the more complicated DA method" (line 195) and the authors therefore recommend against using DA (line 196). However, I feel the authors are overstating the results: differences in improved performance between AR (10%) and DA (8%) are relatively small, as also seen in Fig. 3 where the DA lines (red) and the AR lines (orange) are close.

We aren't sure how to address this. This difference is significant.

2. On line 51 it is stated that "the purpose of this paper is to provide insight into trade-offs between DA and AR". I feel the paper doesn't entirely deliver on this. Yes, the two techniques are compared across a large number of basins, but the reader doesn't get a clear sense when to use which technique. Appendix F contains a regression analysis in this direction but concludes that "we were generally unable to predict differences between the NSE scores of DA and AR". Closer inspection by a human however may lead to some insights. E.g. it could be interesting to look in more detail at extreme cases: ones where AR significantly beats DA, and vice versa. For example figure 2 shows dots in the south that are green (good) for AR and purple (bad) for DA, and vice versa.

We added nuance to much of the discussion about the comparison between these approaches. However, the bottom line is that autoregression works better in that it is generally more accurate, easier to implement, and less computationally expensive. The reviewer is correct that there is scatter in the benchmarking scatterplot. Unfortunately, as discussed in the appendix, this scatter is not predictable (as far as we have been able to discover). Because the scatter is not predictable, there is no way that we know of to determine where or when using DA might be useful. This is discussed explicitly in the paper.

3. Related to the previous comments, I think the paper in general would benefit from a more balanced and nuanced discussion of the usefulness of both techniques, i.e. the trade-offs. For example, on line 52 the authors claim that "AR is easier to implement than DA". One

could also argue that DA is "easier", or at least more modular, since it does not require changes to the model. Similarly, on line 191 the authors state that "we have no reason to suspect that other DA methods might perform better than variational DA". Without additional explanation or insights, this statement is not supported by the results in the paper. Given the wide range of DA approaches and implementations, it is not clear why this statement would hold. See also comment 5.

See previous response. Although DA in this context is easier to implement than for traditional (conceptual and process-based) hydrology models, AR is still much easier to implement and computationally less expensive during inference.

4. Metrics, section 2.3: please specify what kind of forecasts you are evaluating, are these nowcasts?

Yes, you are correct. We are doing nowcasting. We added this to the methods section.

5. Methodology: results of DA typically strongly depend on how error parameters are set. Details on this aspect are provided in the appendices. We have error covariances B and R in eq.B5, which translate to alpha parameters in eq. C1. These alpha parameters are tuned during an independent validation period, with values reported in Table E1. We see that the tuned value of alpha_c (how much we trust/weight the trained model) is zero, and that alpha_y (how much we trust the real-time data) is fixed at a value of 1. If I understand it correctly, setting instead alpha_c=1 and alpha_y=0 in eq.C1 would fall back to the benchmark simulation model, i.e. not using real-time data. Why then not also tune alpha_y? Or tune some weight w=[0,1] with alpha_c=w and alpha_y=1-w? That way the DA model includes the simulation model as a special case and should never perform worse. The current results sometimes (Figures G1 and G3) show worse performance for DA than for the benchmark simulation model. Also, are the alpha parameters the same for all basins? Why not estimate separate values for each basin?

alpha_y is set to be 1-alpha_c1. We clarified this in the revisions (last paragraph of Appendix E).

It's a good point that separate hypertuning per basin is possible (even with a single, shared model). In an experiment that is not reported, we estimated separate hyperparameters for inference (DA) in several individual basins, and we also tested directing optimizer gradients toward inputs (both static and dynamic) as well as cell and hidden states in the LSTM separately per basin. There were some interesting results from this experiment, for example in some of the basins, best results came from updating only the static catchment attributes, which indicates that in these basins the catchment attribute data might be poor (this might potentially be a way to identify data errors). But the results doing this per basin vs. in batch did not change qualitatively (AR > DA on average). Coupled with the fact that AR is cheaper (does not require separate hypertuning), and the fact that in operational scenario there are often tens of thousands of gauged basins, which would make basin-specific hypertuning

very expensive, our thinking was that the set of experiments reported in this paper are the most meaningful.

6. Appendix B describes variational DA and its application to LSTM. I think the math needs to be 'cleaned up' a bit for clarity:

We fixed all of the notation issues mentioned in the following comments.

-loss function L is written as function of model inputs x and outputs y, L(x, y), while loss is typically a function of model outputs y and corresponding observations. Where the model output depends on the unknown parameters or states for which derivatives are computed.

-Eqs. B13-B15: I don't think the gradient chains are correct, since they assume h[t] is independent of previous time slices given c[t], while the model equations B6-B11 show that there is an additional 'path' from h[t-1] to h[t]. I understand the appendix is meant to give the reader a general sense of what is happening, but you might as well write it down more correctly to avoid confusion.

-Eq. B14: the derivative on the left should be with respect to c_l

-Eq. B15: on the right we should have x and y from t-s to t instead of from 0 to t? And on the left derivative with respect to c_l[t-s], and x[t-s:t] instead of x[0:t]?

-I found it confusing that Eq. C1 switches to [t, t+s] from [t-s, t] in Eq. B15.

7. Eq. 1: what is epsilon?

8. Eq. 1: don't you want to divide by N here? Otherwise NSE values increase with N...?

9. Line 84: "is reproduced"

10. Line 199: at the time of this review, no code was provided in the linked github repository

Yes, I believe we mentioned this when we uploaded the manuscript. The repository is staged but has not been made public yet. Our intent is to make it public once we are comfortable that there are no major errors in the paper. We will do so before submitting a revised manuscript.