Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-515-RC2, 2021

review of hess-2021-515

Anonymous Referee #2

Referee comment on "Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks" by Grey S. Nearing et al., Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2021-515-RC2, 2021

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.

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.

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.

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

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?

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

-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