Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-511-RC2, 2021

Comment on hess-2021-511

Anonymous Referee #2

Referee comment on "How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?" by Reyhaneh Hashemi et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-511-RC2, 2021

Summary of Review:

This paper addresses two research questions related to the use of LSTMs for rainfall-runoff modeling: (1) Does appropriate sequence length depend on hydrological regime, and (2) should LSTM training be done on hydrologically similar basins?

To state my opinion up front, I have run similar experiments (unpublished) and found results that are qualitatively different than what are reported here. There are several technical issues in this paper (overall, the methodology is not appropriate for testing the stated hypotheses), and it might be worth addressing those before we look carefully at the results.

My overall recommendation is to revise the experiment as suggested in one of the comments below. The experimental design that is appropriate to test the (two) hypotheses outlined here is very simple (but somewhat computationally expensive). If the authors were to find similar results using a more appropriate experiment, this would be an interesting study.

Comments:

Hyperparameter tuning was done on LSTMs trained on individual basins. LSTMs trained on individual basins behave fundamentally differently than LSTMs trained on multiple basins, which means that lessons learned from hypertuning on individual basins do not translate to multiple-basin models. Additionally, 15 catchments is not enough for robust hypertuning – we would need to perform hyperparameter tuning on the full (evaluation) dataset (although see a later comment – the experimental design needs to be changed

fundamentally). Also, notice that the only portions of the "hypertuning" that were actually used for the other experiments in this paper were (1) discarding the S2 model architecture, and (2) batch size.

There is strong relationship between the dimension of the cell state and the sequence length, and also between the cell state dimension and the ability of the model to generalize (Kratzert et al 2019 shows how the model uses the cell state to map catchment similarity). This parameter was not included in the hyperparameter tuning, and it was also not considered in the experimental design. 64 cell states is smaller than used by most of the previously published work. The hypotheses that are tested here are about the ability of the model to generalize and about memory timescales, both of which are directly controlled by the cell state (more cell states means more ability to have different memory timescales for different hydrological regimes).

It would be interesting (and useful) to know whether there is value in clustering catchments prior to training models, and if so whether we could find correlations between different hyperparameters (e.g., sequence length, cell state dimension) and hydrological regime (the former is a more interesting question than the latter, in my opinion). **The way to test this is simple – you separately (and fully) hypertune each model.** For example, if you want to test the clustering strategy described in lines 120-125, you would hypertune models separately for each catchment group (considering all of the important LSTM hyperparameters), and as a benchmark you would hypertune a model for all of the catchments combined. Then the results would be directly comparable. After that, you could look at whether there was any relationship between hydrological regime and the "optimal" (hypertuning is never actually optimal) sequence length for that cluster. If you really wanted to train single-basin models (which I suggest you should not do), then these need to be separately (and fully) hypertuned for each basin.

I wonder why we are training local models. There is no situation where we would ever use a model trained on a single catchment for any real-world purpose. Additionally, the behavior of the LSTM is fundamentally and qualitatively different when trained on one catchment vs. many, which means that we cannot learn anything general or useful from locally trained models. If there was a specific hypothesis that we wanted to test that required training local models, then this might make sense, but I do not believe this is the case here – we could ask the question about appropriate sequence length on hydrologically grouped models, and asking the question this way would give us a more useful answer. Just a note: Kratzert et al. (in all papers after their 2018 paper) trained single-basin models only to make the point that this is not an appropriate thing to do.

Minor Comments:

The S2 architecture (stacked LSTMs) is interesting, but not related to either of the hypotheses of the study. What was the motivation for testing this and how does it relate to the questions that were motivated in the introduction? I'm not saying to remove it, just give us some reasoning or motivation. Also, when the "complexity" of this model is discussed, you might give us the number of free parameters so that we can get a sense of

what the differences are.

Line 192: *"The validation set is intended to be used for finding the best weights and biases during training and control overfitting."* I think this is just a typo. Validation data is used to help find the best *hyperparameters* and control overfitting (it is explicitly *not* used to help tune weights and biases, except through early stopping).

Line 201: *"What remains constitutes the train period (P1) the length of which varies between 1 year to 40 years in the FR sample."* It is a little concerning to have different sized training data records per catchment, especially if some catchments only have 1 year of training data. This is *especially* problematic if we are looking at differences between what data is required to train in different types of catchments.

In line 180 is reads like you are doing sequence-to-one prediction, however in line 259 you say that you are using a patience of 50 epochs with a maximum of 500 epochs. Typically you only need this many epochs if you are doing sequence-to-sequence training. Regardless, the number of epochs used by previous studies was in the range of 20-50. Have you found that more epochs help (we looked at this carefully in previous studies), or is there something else about your model that is different from previously published work?

Line 291: This is a pretty small list of catchment attributes. Given that catchment attributes are available globally (e.g., HydroAtlas), and this will directly influence the generalizability of a model, why did we use such a limited set of attributes here?

In general, naming experiments with non-descriptive names like R1, R2. P1, etc. makes the paper more difficult to read than is necessary. This means that the reader must always refer back to the text in order to understand each figure. This can be solved simply by naming each of the models/experiments/datasets with descriptive names.