Link to Review: PDF

Comments/Text of Anonymous Referee 1 (AR1) posted in blue, and our answers are posted in black.

This very interesting paper of Kratzert, et al. compares the quality of the predictions of various hydrological models with three variants of the Long Short-Term Memory(LSTM) deep learning network. One of these variants, the novel EA-LSTM, is trained using meteorological data and catchment similarities as an additional input and is analysed in detail highlighting the superiority of such a network. In general the paper is very well written and it is worth to be published after some minor changes. Some comments:

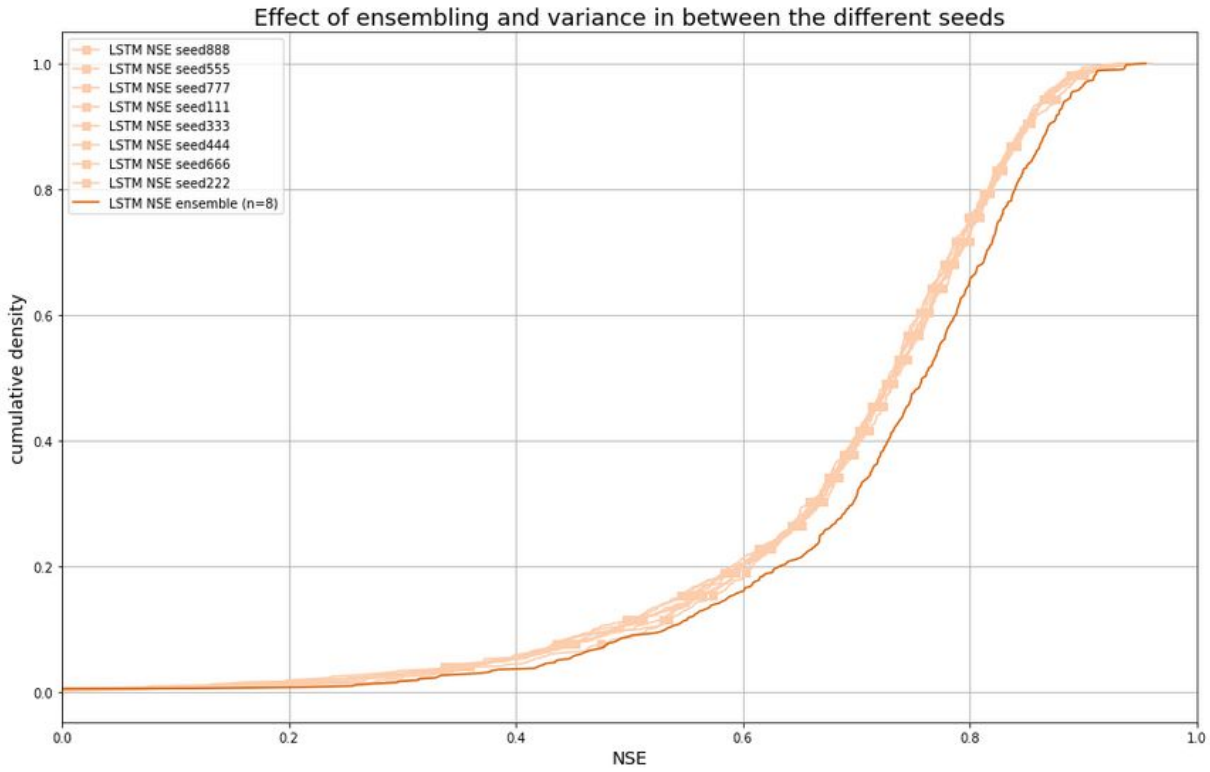
1. Maybe you could explain better the differences of the analysis of the single model and the ensemble mean approach. On page 13, lines 317-320 you write: “To assess statistical significance for single models, the mean basin performance (e.g. mean NSE per basin and across all seeds) between two different model settings was compared between different model configurations.“ What’s the difference between model settingsand configuration? If I understood it correctly the difference in the verification of the single models and the ensemble mean is: Single model: From 8 ensemble model runs, you get 8 different predictions and you calculate the verification measures (e.g. NS values) for each of it and take the average (+/- Std? in Table 2); whereas in the Ensemble mean approach, for example this measure is calculated taking the mean of the 8 predictions?

Thank you. We will rewrite this section of the description of methods (lines 317ff) to more clearly describe the ensemble approach and how the statistics of the single model are calculated.

2. For clarity reasons I would not include the single model outcome in Figure 3, because this a random outcome and would look different for each ensemble run.

We don’t agree with the reviewer on this point because we think it helps to underscore the potential of ensembling. As shown in Table 2 and Table 3 in the manuscript and the figure below, the overall CDF of the single models does have little variation between random seeds, especially in comparison to the benefit of ensembling. Therefore, we would like to keep these curves in the figure, since it helps to visualize the ensembling benefit and show by how much

the CDF can be improved.



3. Nice to have the significance reported, which is most often not shown. Although the precision of these p-values is extremely high and the differences are probably rather neglectable caused by noise.

We agree with the reviewer that reporting significance measures is important.

4. Regarding the modified NSE. Wouldn't it be easier to normalize the streamflow data (e.g. using the BoxCox transformation)? So you don't have to event a new measureand adding a constant in order to achieve stable results.

The loss function we use isn't really new. It's just the basin-average NSE. The reason this is a little unusual in Hydrology is because we rarely (if ever) calibrate a single model to multiple basins. But this is standard practice in machine learning - where the overall loss function is the average loss over many samples. The only alternative is a single loss function calculated over concatenated data from multiple samples, which doesn't work well (or at least not as straightforward) for stochastic gradient descent, which randomizes and sub-samples the training samples.

Moreover, the reason we did not transform the data before training is because this affects model performance (besides normalizing to zero mean, unit variance). For example, an exponential-type transform like Box-Cox will generally under-emphasize peak flows (if the

box-cox exponent is <1). We did try calibrating to log-transformed data, in an effort to normalize the streamflow data, but this does not work as well as using the natural streamflow data. The goal is to train to the non-transformed target data, but use a loss function that does not overemphasize any particular training sample (i.e., any particular or individual basin does not have out-sized influence on the training procedure).

5. Looking at the results, I would conclude that the EA-LSTM is very interesting fort his analysis, but for practical applications the LSTM with the coupled meteo data and catchment attributes is even more efficient and is less complex. That's why I would like to see the results of this model also in Figure 4, 5 and Table 3.

We will modify Figure 4, Figure 5 and Table 3 to include the results of the standard LSTM.

6. Are the catchment attributes kept static for all days of the year? For example the monthly mean of leaf area index could be easily varied depending on the month of the year?

Yes, the catchment attributes are static in this study. We are currently working on making these dynamic, including vegetation and climate indexes, soil moisture, snow cover from remote sensing, etc. This is however not a trivial extension and we do believe that the idea is worth being studied separately. Furthermore, for the sake of repeatability, we wanted to stick entirely to the CAMELS data set, which only includes static catchment attributes. Right now, in this paper, we are using long-term catchment attributes as indicators of differences between catchments (regional heterogeneity among catchments), not for assessing nonstationary catchment behaviors.

7. I would suggest to delete the UMAP analysis, since the method is not explained and the results are a bit confusing.

On this point we don't agree with the reviewer and would like to keep this section. We see this as an interesting addition to the introduction of the new EA-LSTM and the benchmarking results. Specifically, we are using this analysis to illustrate the fact that the embedding layer (our static LSTM input gate) can 'learn' about catchment diversity in a physically meaningful way. This is a (fairly simple) form of explainable AI, and one of the goals of this paper is to work toward that larger objective. The analysis of the embedding layer is important as an example of this larger purpose, and the UMAP analysis in particular is necessary for a reduced-dimension (i.e., graphical) analysis.

Although our paper is mainly intended as a modeling paper and we see the introduction of the EA-LSTM and benchmarking against various hydrological models as our main contributions, we think keeping Section 3.4 has the following benefits:

- It provides at least some feeling about what happened in the embedding layer of the input gate in the EA-LSTM (namely grouping of basins that match our expert knowledge).

- As such, it helps in our opinion to gain trust into LSTMs which are widely considered as black-box model.
- It shows possible analysis that are possible with the EA-LSTM in general and potentially open the door for many follow-up studies in the future. Such studies could either concentrate more on the interpretability of LSTM-based models or try to extract new hydrological understanding from the learned groupings.

We will extend the manuscript to include a short description of the UMAP method, however, would avoid a lengthy discussion on the method and point the interested reader to the official UMAP publication. In our manuscript it is simply take as (state-of-the-art) dimension reduction technique.