Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/hess-2022-245-RC3, 2022

Comment on hess-2022-245

Anonymous Referee #2

Referee comment on "The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment" by Dapeng Feng et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-245-RC3, 2022

This study analyzes the ability of deep learning, and physics-informed learning models to make predictions in regions that are outside of the training set. This is an interesting problem, particularly in testing the limits of learning-based models to make predictions in conditions that are outside of the training set. This paper is a valuable contribution to the hydrological modeling literature, and would like to see it published. While there are some wording issues (listed below), and some issues (also listed below) that I would like to see addressed. In particular, there are a couple of points on the training procedure for the LSTM, which limit its performance, ensembling randomly initialized models and including multiple precipitation products are known to boost the LSTM performance in other experiments, it would be good to check if that would have the same result for ungauged regions. In terms of presenting results, since this paper is about region specific (as in regions held out of the training set) models, I would hope to see region specific results, which are mentioned in the text (in terms of model parameters), but results are not quantified nor plotted.

Line 18: Is “PUR” an acronym for “Prediction in Ungauged Region”, or is it a general term for “regionally held out basins”?

Lines 148-149: Can you please clarify the training periods for all the models. You mention that “Each training instance had two years’ worth of meteorological forcings, but the first year was used as a warmup period so the loss was only calculated on the subsequent one year of simulation”, this reads to me that your models were trained on just on year of data. This isn’t the case, is it? I imaging that, in the training process you cycle through many more years, you just train the model with batches of these individual year? Oh. I read on line 245 that “we used only 10 years of training period”. Okay, can you maybe re-word this?

Line 169: Why was Maurer selected? Especially since many studies suggest Daymet is the more informative forcing, including Feng *et al.*, 2022? It is also the case that using a combination of the three forcing products from CAMELS results in improved model performance (Kratzert et al., 2021), can you expand on your decision in the context of using multiple precipitation sources?

There should be a direct link to the analysis done for this paper. I browsed around the HydroDL github repository, and it was not clear to me where I should look for the code that was used for these particular experiments. NEVERMIND ABOUT THIS. I NOW SEE THE AUTHOR'S RESONSE TO ANOTHER REVIEWER.

The issue of ungauged regions is not particularly relevant to the United States (U.S.), but I do see the value of using the U.S. gauged basins for this experiment. Other groups (Le *et al.*, 2022) have done ungauged region experiments outside the U.S., and this could be a bit more compelling. Perhaps this is worth some discussion in the paper?

Line 245: You mention that there was no ensembling of models trained from random initialization. But then go on to say that you used the same settings as Kratzert et al., 2019, but they used ensembles of 10 models trained with random initializations. From their paper:
"Because of this, the LSTM-type models give better predictions when used as an ensemble. It is not necessarily the case that if one particular LSTM model performs poorly in one catch-ment that a different LSTM trained one exactly the same data will also perform poorly." This is generally an accepted practice when using deep learning models. Can you explain further why you decided not to use model ensembles?

Line 308: Breaking down Figure 6 by region would add a lot more value to the results. This would be super valuable for understanding some of the regional trends in model performance in general, particularly in Regions 4 and 5.

REFERENCES:

Le, M.-H., Kim, H., Adam, S., Do, H. X., Beling, P., and Lakshmi, V.: Streamflow Estimation in Ungauged Regions using Machine Learning: Quantifying Uncertainties in Geographic Extrapolation, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2022-320, in review, 2022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions 540 in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344–11354. https://doi.org/10/gg4ck8

Kratzert et al., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling.
https://doi.org/10.5194/hess-25-2685-2021