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.

Thanks a lot for the professional and to-the-point suggestions. We replied in detail to each comment as below.

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

PUR is an acronym for "Prediction in Ungauged Region" which means holding out large continuous regions for spatial extrapolation tests as proposed in Feng et al., 2021. We have modified the original statement as *"test in regionally held out basins, or Prediction in ungauged regions (PUR), representing a highly data-scarce scenario"* to avoid confusion.

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. Geophysical Research Letters, 48(14), e2021GL092999.

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?

Thanks for reminding us of this potential confusion. For the training period of spatial generalization tests (PUB and PUR, train and test in different basins), we originally stated in line 190 and now modified a little as *"The study period was from October 1, 1989, to September 30, 1999. These spatial generalization tests were trained and tested in the same time period but for different basins".* We use data from this ten-year period to train the models. As for the statement in line 148 as *"Each training instance had two years' worth of* *meteorological forcings…"*, this is talking about how we form a minibatch consisting of short training instances from the whole 10-year period. Deep learning models need to be trained on minibatch data that are formed by training instances of time series. Here we mean for each training instance, we randomly select two years meteorological series from the whole training period in one basin. The first year was only used to warm up the state variables while the second year was where the loss function was calculated. The training instances from randomly sampled 100 basins will form one minibatch. Then the model was forwarded on this minibatch to calculate the loss function and do gradient descent to update the weights. We will modify the original statements about the training processes to avoid confusion.

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?

This is a good point. Actually, we hesitate for a while to show the results of which forcing between the Daymet and Maurer. We finally choose Maurer forcing because we want to be consistent with traditional regionalized models like MPR+mHM used as comparison in Figure 3 and 5 that use Maurer forcing. We agree that different forcings will impact the modeling results and the key point is the used forcing should be consistent for performance comparison. In the revised ms, we will also show the performance of Daymet forcings and add related discussions about this.

We agree that using multiple forcings can be beneficial when pursuing the best performance for DL models, and DL models also have advantages to efficiently integrate multi-source data compared with physical models. Using multiple forcings would be more difficult for differentiable models which maintain mass balance (not impossible, but will require some work, and we will get to it in the future as our method is so new). We will discuss this in the revised manuscript. This is also discussed in our previous PUR study in Feng et al., 2021 that DL models can easily make use of auxiliary data. However, the key point here is to make all experiments comparable (keep the same setup) rather than gain the best performance. It's worth emphasizing that we don't aim at invalidating the power of LSTM models, and LSTM is part of the parameterization pipeline for the differentiable models. Instead, we want to show how well the differentiable models can generalize in space with the previous regionalized methods providing context. We will add discussions in the ms to further clarify this point.

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.

Yes. For anyone who has not seen it, the codes were released at zenodo with this link to access: https://doi.org/10.5281/zenodo.7091334

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?

We agree that PUR is very important for global hydrologic modeling but the uncertainty has not gained enough attention. There should be extensive work on global hydrologic modeling using differentiable models in the future. The final aim would be better modeling the global ungauged regions and we think the most difficult problem is the cross-continent prediction. We will add some discussions about this global modeling topic to motivate future research.

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?

Thanks for pointing out ensemble modeling. As we stated above, the key point is we should make the setup of all models and experiments comparable to examine the spatial extrapolation capabilities. With our LSTM, we repeated the same setup (best performance setup) as the ensemble LSTM job in Kratzert et al., 2019 for benchmarking.

When comparing different models for spatial generalization, we employed a less expensive but consistent setup without using the ensemble simulations. We also keep all the modes using the same random seed for initialization. The PUR and PUB tests already use cross validation with 7 and 10 models running for each experiment which can somehow reflect ensemble characteristics as used in Beck et al., 2020. Cross-validation multiplied by ensemble will be too time-consuming. In addition, we wish to do a more thorough evaluation of various ensemble methods in the future, e.g., we can even make an ensemble by perturbing the inputs to neural networks named input-selection ensemble as shown in Feng et al., 2021. It is indeed possible that a particular ensemble method would give advantages to certain methods under different conditions, and we think this complication would need a lot more computational and experimental time to examine. We will acknowledge this as a limitation and a future work plan item but we hope the reviewer could see that there are many things to do on our plate right now with the new differentiable models and their optimal setup is still very much fluid.

In the caption of Figure 3 we gave explanations as "*The PUB was run in a less computationally-expensive training experiment to be comparable to other models and also to reduce computational demand: we used only 10 years of training period, did not use an ensemble, and used a lower k-fold. When we ran the experiments using the same setting as Kratzert et al. (2019), our LSTM was able to match the PUB performance in their work (Feng et al., 2021).*" We will add more clarifications and discussions regarding these two aspects (ensemble & multi-forcing) in the revised manuscript.

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. Geophysical Research Letters, 48(14), e2021GL092999.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17), e2019JD031485.

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.

Thanks for this good idea and we will add a figure showing the NSE/KGE metric in all the seven PUR regions as well as the related discussions in the revised ms.

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