Response to RC#1

Title: Streamflow Estimation in Ungauged Regions using Machine Learning: Quantifying Uncertainties in Geographic Extrapolation

General:

[1] This paper attempts to make predictions of monthly averaged streamflow in data scarce regions with machine learning models that were trained in data rich regions. They test their predictions with different permutations of training regions. As expected, the models perform better with different climates and catchments attributes in the training set. Interestingly, however, thew results suggest that models trained in North and South America are more reliable than models trained in Europe. They also find, as expected, that extreme gradient boost outperforms support vector machine and random forest. The paper is written fairly well, with exceptions noted below, and provides additional support for the well established conclusion that machine learning models trained on diverse data sets can be useful outside the basins which they are trained. This paper expands that conclusion by transferring the learned models to entirely new regions, in particular to data sparse regions, which is important, as the authors point out.

Response: Thank you for your positive comments and valuable suggestions which will help us improv our paper's quality. Below we have answered your concerns point by point.

[2] It was not clear to me if these models were forward looking or backward. I am not entirely sure how useful a monthly average streamflow prediction is in practice, especially if the forcings which drive the prediction are aggregated over that particular month, which would have the prediction a backward estimate. If, however, the forcings are aggregates from the previous month, then this is valuable to water resources management. I ask the authors to make this clarification in their data and methodology sections.

Response: Our current analysis is backward looking. We understand the limitation you mention and will address your concern in the revised manuscript. That being said, our study is unique in that few attempts have been made to examine the transfer learning concept in a real-world context in the field of hydrology. We tested the pre-trained ML models' streamflow prediction using the assumption of no prior knowledge of the new streamflow prediction domains. These new streamflow prediction domains are located in data-poor regions (Central Asia and South Africa) where water security is a long-term problem. In reviewing the case study, we saw that backward analysis would be particularly useful to us. From the GSIM datasets, aggregated climatological streamflow is the most feasible way to obtain more sample points for our analysis. More importantly, in transboundary river territories where obtaining in-situ streamflow can be challenging, the climatological streamflow is beneficial for long-term water resources planning. These averaged climatological streamflow indices are considered to be the minimum number required for water resource planning in a transboundary river basin in Northeast Vietnam (NAWAPI, 2018).

References:

NAWAPI, 2018. Bang Giang - Ky Cung Water Resources Planning Project, Water Resources Assessment Report (in Vietnamese), Ministry of Natural Resources and Environment, Hanoi.

[3] This paper omits non-machine learning models from the study because they are harder to set up. And unfortunately, there is no benchmark model/s presented. I believe that this could potentially draw criticism. I do fully understand the need for easy-to-use models in some situations. I would encourage the authors to

rethink their framing of the model selection in the introduction and conclusion. Perhaps it would be good to make a case for the benefits of easy-to-use models, and make a case that these shallow learning models are suitable for monthly averaged streamflow over the state-of-the-art LSTM mode, which has been shown again and again to outperform other streamflow models, even when trained out of sample.

Response:

We thank the reviewer for their suggestions. In the revision, we plan to compare our reported ML models with an state-of-the-art land surface model (i.e., Noah-MP4.0.1) that uses four different routing schemes: "free drainage", "groundwater", "TOPMODEL" and "BATS". In addition, we will also consider developing LSTM and GRU deep learning models to assess whether they will introduce substantially different outcomes.

Abstract:

[4] Line 21 has double periods.

Response: Thank you for noting this typo. We will revise accordingly.

Introduction:

[5] Lines 38 and 39 claim about stream gauges being the most accurate way to measure streamflow is vague and trivial. Are you making a distinction between remote sensing and in situ measurements? There are many methods of gauging a stream, some more accurate than others. I'm not sure what is the purpose of the sentence, remove or clarify.

Response: Thank you for this note. We will revise accordingly.

[6] Lines 78 and 79: If there is a good argument that ML is not **the** most promising approach, I'd like to see a citation. Otherwise just state it directly as "machine learning models are arguably one of the most promising approaches"

Response: Thank your for your valuable suggestions. We will revise accordingly.

[7] Line 86: I'm not sure it is obvious what at "traditional" hydrological model is.

Response: by "traditional" hydrological model, we mean a physical-based hydrological model such as SAC-SMA, VIC, mHM, and the National Water Model. We will revise these lines accordingly.

[8] Line 107-108: I assumed your hypothesis was about ML model's ability to transfer learning from one region to another, but here you claim that you use ML models because they are easier to set up?

Response: Thank you for your interesting comments. By “easier to set up,” we were comparing our ML models with physical-based hydrological models. Setting up a physical-based model for individual catchments requires considerable processing time and assumptions of soil type, topography, and physics. Such a model is challenging to work with in handling large pool catchments. We will rephrase this sentence to clarify it for readers.

[9] Lines 108-109: I think the last sentence of this paragraph is fragmented. What kind of water resources prediction? In what context are the water resources secure or insecure?

Response: Our results are helpful for long-term water resource planning as our model outcomes (climatological streamflow) can provide general information for water resource managers in regions with limited or no access to in-situ networks. This information is probably essential in many transboundary rivers. We will revise this paragraph for the sake of clarification.

Data:

[10] Line 127: What is the rational for removing values greater than 2,000 cms?

Response: Monthly streamflow values greater than 2000 cms do not seem feasible, so we excluded them to avoid noisy data in the post-processing datasets (Ghiggi et al., 2019).

References:

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth System Science Data, 11, 1655-1674, 2019.

[11] Line 140: Can you make it clear if your model is making a forward or backward prediction? Is your monthly forcing aggregates from the same month in which your monthly averaged streamflow comes from?

Response: In its current state, this work is a backward prediction forcing of aggregated values from the same month for P and T.

[12] Figure 1: What unit is catchment density?

Response: It is a non-dimensional unit reflecting the relative degree of dense catchments over a specific region.

Methodology:

[13] Lines 207-209: This wording is a little confusing. Can you rephrase to make it clear that the validation set was used to tune the hyper-parameters? Meaning, your training set is used to get the model weights, and then you check the quality of those weights by calculating an error on the validation set, then modify a hyper-parameter and train again, then check the quality of the new weights on the validation set. And to be clear, you do not calculate any error on the test set until the hyper-parameters have been chosen, right?

Response: Your understanding is correct. We will revise accordingly.

[14] Table 3: Consider moving the regions into the table header, instead of as a note.

Response: Thank you for this note. We will revise accordingly.

Results and Discussion:

[15] Line 236: "The local-based models also served as benchmark models" This should be moved to the methods section.

Response: Thank you for this note. We will revise accordingly.

Limitations and further studies:

[16] Line 326: In parentheses you have "daily or monthly", but I think you meant "daily or hourly"

Response: Thank you for this note. We will revise accordingly.

Conclusions

[17] Line 334-335: "ML algorithms to quickly test our hypothesis since ML algorithms could be easier to set up than traditional hydrological models." I think this is a bad reason to us ML. There is no use doing a study with one tool instead of another simply because it is easier.

Response: We greatly appreciate your point. In our experimental setting, we would use a straightforward term to describe the advantage of ML over traditional hydrological models [or physical-based models]. We will revise our word choice to better explain our model selection.

[18] Line 351: double periods.

Response: Thank you for this note. We will revise accordingly.