Response to RC#2

[1] This paper investigated the monthly streamflow prediction in ungauged regions using Machine Learning (ML) based methods. The authors compared three ML methods in global basins with two large regions as data-poor targets. The overall structure of this ms is clear to follow and the topic is intriguing to me. I have some comments as shown below on better clarifying the methodology and performing more profound discussions on the results to safely draw the conclusion. Hopefully these suggestions can help to improve the quality of this study.

Thank you for your positive comments. We have answered your questions and comments point by point below. We will incorporate your suggestions to improve our manuscript's clarity.

[2] Introduction: The authors did a good job here with a comprehensive review on the present studies and I enjoyed reading this part.

Thank you for your kind comments.

Methodology:

[3] To my knowledge, the present cutting-edge ML applications in streamflow prediction mainly focus on daily prediction with deep learning models like LSTM which show superior performance over other models as shown by several studies already cited in this ms. The advantage of DL models over traditional ML was not only shown in hydrology but also in many other fields. I feel the authors may discuss more on the motivation of their choices on monthly prediction and model selection with traditional ML methods.

This study focuses on data-centric ML rather than method-centric ML. Our significant contribution is proposing a new dataset for testing Prediction in Ungauged Regions (PUR) in a real-world case study. Our study is unique in that we have attempted to predict streamflow across continents with diverse climatic and catchment characteristics. Previous works with LSTM often used datasets which required minimum effort to process and which focused on a single continent. Since data availability in our experiments varies from place to place, we selected climatological monthly predictions to maximize the number of available catchments. Our focus is intended to solve a spatial prediction (predicting streamflow in different geographic/ continent regions). At the same time, LSTM is favored for temporal prediction; thus we did not consider LSTM in this manuscript. Finally, using this method, our sample data for the training set could consist of fewer datasets than would normally be favorable for LSTM.

[4] Better clarification on the framework and experiment design is needed to help readers easily understand the method section. I am quite confused about the meaning of "100" mentioned in line 219 and throughout the ms. Does this mean a 100-fold cross validation to cover all the available data? If so, there would be no basin overlapping for each testing but how the 100-simulation range comes then? I also didn't understand how the training, validation and testing dataset were formed with limited details given.

Your understanding is correct. The 100-fold cross-validation was used to assess the uncertainty of each ML algorithm. For one input set, we randomly selected samples with replacements. We repeated this process 100 times, so one catchment could be in the testing set in one fold but in the training set or cross-validation set in another fold. We did not fix the testing set since we wanted to examine the sensitivity of the ML algorithms with different spatial data combinations.

[5] How do you organize and divide data in the time dimension? The streamflow prediction is a time dynamic problem and I see the authors use data across multiple decades, however I only find the results reported for 12 individual months without time continuous information given.

We aggregated monthly data to climatological monthly data. This is why you see our results reported for 12 individual months rather than continuously. As we mentioned previously, our rationale here was to observe streamflow data availability; climatological streamflow is the best way to obtain the maximum number of catchments.

[6] If I understand correctly, the authors train individual models for different months. I am just curious how this choice was made and how the model would behave with one model trained on data from all months instead, especially given the power of ML models handling big data.

Since different months have different seasonality cycles, we believe that having a separate model for each month would be better than having one model for all months. We did test one model for all months, but we saw that this model did not perform well as the separate models for each month.

Results:

[7] Reading through the result section, I hope the authors can do a more profound analysis and discussion on their results, not limited to simply describing the figures. The present figures are kind of redundant to me especially without many discussions involved. You may consider removing some unnecessary ones.

Response: We appreciate this comment. In the revised manuscript, we will enhance our results section by incorporating your suggestions.

[8] For the PUR performance evaluation, the authors need to clarify more about the absolute performance in target regions, not only the performance difference from the local models. It's quite intuitive to get worse PUR performance compared with local models, but the readers care more about the direct evaluation, like how will ML models behave and can we get functional models for predictions in ungauged regions? Looking at Figure 8, I feel the absolute PUR performances are mostly close to KGE value of 0.0 (y axis starting at -2.0 can be somehow misleading to readers), which implies unsatisfactory performance for a functional model.

Response: Thank you for your interesting note. We will add a direct values comparison to our revised manuscript. We observed positive PUR performance (KGE >0) over the course of several months. We used -2.0 to show the full possible ranges of KGE performances from the source-based models demonstrating model uncertainties in predicting streamflow at the target region. In the revised manuscript, we will provide a more detailed explanation to support our assertion that functional models can perform quite well.

[9] It's quite interesting and also surprising to me for the statement of line 290 that including more training data (EX7 here) leads to lower performance. I hope the authors can have more investigation and discussions on this point, which could be quite controversial given the common agreement that ML models usually benefit from bigger data. Thinking about this, I feel it may depend on different scenarios, such as different types of models used with different capacities to handle large data, and how you train and evaluate the model - the model with more input data may not get optimized which leads to underfitting. Taking one example, for experiments EX1-EX7, the optimized hyperparameters can be different with varying training data availability, and a fair comparison should be built on the optimized conditions of different models.

Response: The results presented here are for the pre-trained ML models in the entirely new study domain (the target). Therefore, our assertion does not conflict with the common agreement that ML models perform better with more training datasets (Figure 4). Our message is that using models forced by the greatest possible number of datasets does not necessarily ensure that those models will perform well. The newly added datasets may even add noise to the models. Specifically, we see that including European catchments may not be a good idea for creating pre-trained models to make reasonable predictions at targets in Central Asia and South Africa. This is probably due to the characteristics of European catchments being so different from those at the target and the fact that learned ML models from these catchments are not beneficial.

[10] I didn't understand the results shown in Figure 3 well. Are these the results on source (gauged) or target (ungauged) basins? Are they reported on the testing data, and if so how did you divide the testing data?

Response: In Figure 3, the results show the testing data for the sources from experiment EX7 (input datasets include S1 – North America, S2 – South America, and S3 – Western Europe). We did not fix the testing data; instead, we randomly divided it into training, cross-validation, and testing sets. One catchment could fall into a training set in one simulation but into a cross-validation or testing set in another simulation. Thus, Figure 3 shows the performance of models trained with different input datasets from EX7. From these results, we gained insight into the uncertainties of ML algorithms as they respond to different input datasets. More importantly, this method allowed us to apply a set of hyper-parameters (associated with different input datasets) from the source (EX7) to the target region. We will revise the manuscripts to help readers better understand our deliverable message in this figure.

Conclusion:

[11] As mentioned in the above comments, I feel the two key points in line 341 and line 343 are kind of contradictory regarding whether more diverse data can lead to better performance or not. The authors should carefully investigate this point before drawing a conclusion here. In addition, as mentioned previously, more analyses on the absolute PUR performance are needed to get the strong conclusion in line 351 that

these models can be capable of solving PUR problems in ungauged regions, especially given the deteriorated performance shown in Figure 8.

Response: We appreciate this comment. In line 341, we discuss the case that unseen data are in similar geographic region as more training data points likely to improve the prediction capacity of the pre-trained model at the same region.

In line 343, we aim at the performances of pre-trained models at an entirely new geographic location (i.e., transfering the model to new region). In this context, the common agreement that pre-trained models with more traning samples will peformed the best is not likely true. In line 351, we agree that the conclusion sounds too optimistic about the performance of the pre-trained models. Specifically, the pre-trained models have performed well in predicting streamflow in several months - but not all months. In our revision, we will revise this paragraph substantially to communicate our findings better and ensure that (any) caveats are presented in a transparent manner.