Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2022-320-RC2, 2022

Comment on hess-2022-320

Anonymous Referee #2

Referee comment on "Utility of Deep Learning and Data-rich Regions in Predicting Monthly Basin-scale Runoff in Ungauged Regions" by Manh-Hung Le et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-320-RC2, 2022

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.

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

Methodology:

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.

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. 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.

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.

Results:

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.

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.

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.

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?

Conclusion:

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.