Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-371-RC1, 2021

Comment on hess-2021-371

Anonymous Referee #1

Referee comment on "Impact of spatial distribution information of rainfall in runoff simulation using deep learning method" by Yang Wang and Hassan A. Karimi, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-371-RC1, 2021

In my eyes this publication sets out to examine an interesting and indeed important research question. Which is: How much information can be leveraged by LSTM based **forecast** models from spatially distributed inputs.

There are many minor issues in the paper that I would like to discuss with the authors, but — as it stands now — the manuscript does not meet HESS standards. It can simply not be judged properly. This states results from very basic choices of setup and exposition, which I'll summarize in three points in the following:

- **Setting:** To me its not clear why the authors chose two specific basins in a forecasting setting? Regarding the former: Current best practise is to train the models on many basins, since single basin training does not yield particularly good results (see for example: Gauch, Mai & Lin; 2021)., The authors do not explain why they move away from these practise. I suspect it has been done to reduce workload, but this is only an inference, since the manuscript does not explain this choice. Regarding the latter: Why was a forecasting setting chosen over a simulation setting. The inclusion of runoff in the features will unavoidably explain away (Wellman & Henrion, 1993) potential influences of the distributed input, since it already integrates over the past. I can see a potential reason in trying to avoid the decrease in performance of a model when using only two — instead of multiple — basins. However, the choice seems to lie in direct contradiction to the goals lined out by the authors (i.e. understanding the influence of spatially distributed inputs).
- **Method**: The setup is unclear. I was not able to understand how the HRUs have been delineated and how the distributed meteorological inputs have been obtained. If the standard CAMELS data is used — as indicated in the code and data availability paragraph — then a lumped input was somehow disaggregated to match the HRUs. How is this done? A naive way would perhaps be to simply weight the meteorological inputs corresponding to the area of the HRUs. If that is the case, the authors would need to show that their new model is not simply better because it has more parameter than the baselines. And, also, how the results would change if a different weightings are used, since it is not a-priori clear why the specific HRU delineation does improve the

result (if it does so). These are all interesting questions. But, as things are currently explained I was not even able to infer how the authors did this and which data really was used. I believe that a clarification here would be of great worth to the readers.

- **Results**: Forecasting (not simulation) models are compared on the basis of two basins with one and half a year of data (arbitrarily chosen from June 6, 2010 to December 23, 2011) each. The resulting outcomes are reported with 5 digits of precision. I suppose this is done to make the models comparable, given that forecast models tend to produce very accurate predictions. However, given the high measurement uncertainties and the potentially large runoff-variability between years this is not possible! There are many other studies (the authors even cite some of them), which report less digits, while evaluating their data-driven approaches on hundreds of basins using multiple years of data. Given the circumstances of the setup, I would suspect that only one or two digits should be reported. And, given the close results I would think that some basic statistical test are necessary, and it also would be good to provide some error bounds related to either how much the results would change with longer/different data and repeated model runs (best both).

There are many other minor points that I'd like to discuss with the authors (for example, I do not see why a calibration period needs to be defined after a training period. Perhaps this is meant to be a validation period?). Before that I would however like to see the manuscript adjusted so that at least these basic points are met and the manuscript can be judged properly.

References:

- Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, *135*, 104926.
- Wellman, M. P., & Henrion, M. (1993). Explaining'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(3), 287-292.