Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2022-113-RC2, 2022

Comment on hess-2022-113

Matteo Giuliani (Referee)

Referee comment on "The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL)" by Juliane Mai et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-113-RC2, 2022

The paper contributes a comprehensive model intercomparison across 13 hydrologic models, including Machine Learning based, conceptual, and physically based models. The analysis is run over the Great Lakes region looking at the model's ability to simulate streamflow, actual evapotranspiration, surface soil moisture, and snow water equivalent. The comparison is performed looking at simulated output aggregated to basin-scale as well as at grid-level, considering temporal and spatial validation. The study is extremely well designed and it provides a solid contribution to the existing literature. The manuscript is also well written and definitely interesting for HESS readership. I only have a few comments that I would recommend addressing before accepting the paper for publication.

1) The model intercomparison reads extremely solid in terms of using consistent data, forcing, etc. as well as in terms of the adopted calibration-validation scheme. Yet, the description of the different models' calibrations in Section 2 seems to introduce quite some variability whose potential implications are not discussed. Although most models have been calibrated using the same algorithm, it is not clear whether the different modeling teams had some guidelines/constraints about the calibration effort to somehow harmonize it across models. Since it does not seem there was any limit on the number of model evaluations run (or on the total time spent) during the calibration, I'm wondering whether some results could be explained by better/worse calibration results. This aspect could also be an interesting finding of the analysis, but to be fair it should be derived by coordinating the calibration efforts. For example, the LSTM model involves 300,000 parameters, and being a data-driven model is by definition more flexible than the other models considered. This LSTM model was calibrated in 2.75 hours; how about the other models? Is the effort of running 300 iterations for calibrating the 9 parameters of the LBRM-CC-lumped model comparable?

2) One of the key assumptions of the analysis is considering only streamflow gauges in low-human impacted watersheds. While the authors clearly motivate this choice, I believe the paper would benefit from some further elaboration around this point given the somehow limited number of "pristine" river basins worldwide, see for example Belletti et al. 2020. Which type of bias we could expect in using these models in a human-impacted

basin? Are these biases consistent across models, or can some categories better capture human inference even if not explicitly described? I believe this type of reasoning could be a good addition to the model discussion, which could be perhaps potentially supported by looking at the model performance in some sampled stations currently excluded from the analysis.

Belletti et al. (2020), More than one million barriers fragment Europe's rivers, Nature, 588, 436-444

3) The temporal validation of the models is based on model simulations over the period 2011-2017, with the models calibrated over 2001-2010. This looks certainly good, but I was then expecting the authors to somehow comment/discuss the role of nonstationary forcing as I expect that data (e.g. temperature) could show some trends over these 17 years. If this is the case, how did you handle such trends? Were the data de-trended or did you use the raw observations? Moreover, what are the authors' recommendations for developing hydrologic models under such evolving conditions? Again, are there any class of models more prone/robust to possible extrapolation biases induced by global warming?

4) Lastly, as the authors probably know the paper is quite lengthy and it does require substantial commitment to get to the end. I think the authors did already a good effort in guiding readers using a good structure and providing summaries of each section, but I would suggest - if feasible - to further shortening the paper in order to facilitate a complete read. I don't have clear recommendations on how to do this; perhaps an idea could be to move the model description of section 2.4 into an appendix keeping only a summary in the main text?