Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-113-RC1, 2022

Comment on hess-2022-113

Anonymous Referee #1

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-RC1, 2022

Mai et al present a thorough model-intercomparison for the Great Lakes region.
The manuscript is very extensive, as is the way that the model intercomparison was managed. The intercomparison is conducted in a very structured manner and clearly was not opportunity-driven; teams had to create a new model-set up to be consistent with underlying data, and perform a new calibration. Also the analysis is very thorough and very honest, fairly comparing the performance of all the models based on different aspects. This is very much appreciated. It also demonstrates how much information can be gained from such a carefully designed experiment: many conclusions on different aspects can be drawn. This does make the manuscript quite long and has the risk that some conclusions might get lost in accompany of all the other conclusions, but the abstract provides a good summary.

I only have a few minor points:

From the methods-section it is not clear whether the LSTM was also trained with geographic data. Later I read it was, perhaps this can already be clarified earlier.

The donor-basin rule is indeed very basic... and as such I am wondering about the value of the space-validation. What does it mean when a model is good at simulating a catchment it hasn't "seen", with parameters based on another catchment? Does that make a model "better"? It could also just be an indication of how sensitive the output of this model is to different forcing / its own parameters, rather than a value-judgement of its performance. But this is just my thought.

It is appreciated that mistakes in the procedures are openly shared, such as about the PET-controlling constant for LBRM-CC calibration. However, there are no consequences related to this point. For instance, it is used as an argument to explain lower performance, but a lack of applying a constraint should actually result in better model performance because during the calibration there was more freedom to fit this parameter (or in equal model performance because the calibration algorithm did end up at the correct spot after all). The implications of this error for comparability are not clear. (same for the other calibration bug with SVS LSS)

It is nice that the majority of the models applied the same calibration algorithm, but all

used slightly different settings. Was this determined based on expert judgement?

Some models were calibrated regionally, other locally. It is unclear why which models where used in one way or the other. I guess because this fits the general philosophy of this model / its common use. Maybe this can be clarified in Ch 2

l. 487-490 (p19) unclear what is meant here.

l. 604 (p23) not very clearly explained.. I guess it also depends on the shape of your pareto front (if it exists at all). It would be nice to see it somehow in a 2D-version (e.g. for two variables only), or a 3D version.

The link to the website is now mentioned quite late. It is a very nice feature, would be nice if it would be mentioned earlier in the text.

In the conclusion it is clearly stated that gridded evaluation might be preferred over basin evaluation (both could demonstrate different results). This is not mentioned as such in the abstract, where only the difference between the two is mentioned.