Dear Matteo,

thanks a lot for your positive feedback. We very much appreciate the time and effort you made to evaluate our manuscript. We happily reply below in detail to your comments. Your comments are formatted *italic*; our replies are highlighted **bold** and ***bold italic***. The line numbers in **red** are referring to the manuscript version you reviewed.

Many thanks,
Juliane Mai and co-authors

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.

Thanks so much for this positive evaluation of our manuscript. We are happy that the work is making a contribution to the existing literature. Thanks again for your time and effort to go through this extensive and long manuscript.

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?*

Thanks for the comment. The models are very different in nature. Some run extremely fast and allow for more model evaluations and even several independent calibration trials from which the best was picked in the end. Other models have runtimes of several hours and can only be calibrated with smaller budgets to be feasible. In majority of cases, these teams relied on a common algorithm based on their individual experiences using it in past studies. We did not enforce the algorithm choice on anyone. We did not want to restrict the models' performance by enforcing same budgets. The task for the experts was to provide the best model setup they can deliver with a given set of inputs. The rest was up to their judgment. We thank the reviewer for the question and will add the following statement to the manuscript (addition highlighted in *italic* font):

line 208 ff: The first group is the Machine-Learning based model which happens to be also the only model with a global setup (Section 2.4.1), the second group is comprised of the seven models that are locally calibrated (Section 2.4.2) and the third group is the five models that followed a regional calibration strategy (Section 2.4.3). *The calibration strategy (local, regional, global) and calibration setup (algorithm, objective, budget) was subject to expert judgment of each modeling team. The main goal of this project was to deliver the best possible model setup under a given set of inputs; the standardization and enforcement of calibration procedures would have limited this significantly due to the wide range of model complexity and runtimes.* The models are briefly described below including a short definition of these three calibration strategies.

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.*

We are sorry that we caused a misunderstanding here. The basins we have picked are indeed NOT all low-human impact; only the basins classified under "Objective 1" are low-human impact (see for example Tab. S15 in the Supplementary Material). The table caption actually states "[...] The objective for each basin is assigned to be 1 if the watershed is of low-human impact while it is assigned to 2 if the gauge station is most downstream to one of the five lakes or the Ottawa River.[...]". All basins that are objective 2 and not objective 1 are considered to be not low-human impact.

In summary:

[station tagged as "objective 1" only]: The watershed is low-human impact and not most downsteam to one of the five lakes or the Ottawa River. There are 66-29=37 such calibration and 33-14=19 such validation stations.

[station tagged as "objective 1" and "objective 2"]: The watershed is low-human impact and most downsteam to one of the five lakes or the Ottawa River. There are 29 such calibration and 14 such validation stations.

[station tagged as "objective 2" only]: The watershed is most downstream to one of the five lakes or the Ottawa River but is not low-human impact. There are 104-29=75 such calibration and 52-14=38 such validation stations.

The numbers of stations in each of the three categories listed above are already given in the caption of figure 1:

Caption Fig. 1: Panel A shows the location of stations used for calibration regarding stream- flow: 66 of them are downstream of a low-human impact watershed (objective 1; large black dots) and 104 stations are most downstream draining into one of the five lakes or the Ottawa River (objective 2; smaller dots with white center). In total, there are 141 stations used for calibration as 29 stations are both low-human impact and most downstream (large black dots with white center; 141 = 66 + 104 - 29). Panel B shows the 71 validation stations of which 33 are low human impact, 52 are most downstream and 14 are both low human impact as well as most downstream (71 = 33 + 52 - 14).

We suggest to make the description of the objectives and the distinction of the three cases more clear in the manuscript (suggested addition highlighted in italic font):

line 441 ff: [...] streamflow gauges need to be either downstream of a low-human impact watershed (objective 1) or most downstream of areas draining into one of the five Great Lakes or into the Ottawa River (objective 2). *If a watershed is most downstream and human impacts are low, the station would hence be classified as both objective 1 and 2.* Objective 1 was defined to give all models – especially the ones without the possibility to account for watershed management rules –

to perform well. Objective 2 was chosen since the ultimate goal of ~~most~~ *many operational* models *in this region* is to estimate the flow into the lakes (or the Ottawa River). *This classification of gauges was a part of the study design that ends up not being evaluated in this paper. This is because each of the modelling teams decided to build their models using all Obj. 1 and Obj. 2 stations treated the same way. As such our results do not distinguish performance differences for these two station types. The information is included here so follow-up studies by our team and others can evaluate this aspect of the results.*

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?*

We indeed know that the calibration period (2000-2010) is a dry period while the validation period (2011-2017) is known to be very wet. This is reported, for example, here: `https://www.lre.usace.army.mil/Portals/69/docs/GreatLakesInfo/docs/UpdateArticles/update206.pdf?ver=2020-07-01-115844-313`**. We suggest adding the following to the manuscript:**

line 440 ff: It is known that the calibration period (2001-2010) is a dry period while the validation period (2011-2017) is known to be very wet [US Army Corps of Engineers: Detroit District, 2020]. This might have an impact on model performances - especially in temporal validation experiments. In this study no specific method has been applied to account for these trends in the meteorologic forcings.

This is certainly something that one might want to take into account while model building and training. We however did not. The very good performance of the LSTM-lumped model in temporal validation (median KGE of 0.82 compared to median KGE of 0.97 in calibration; Fig. 3C vs. Fig. 3A) shows that it might not even be necessary in order to achieve good performance. The larger impact on model performance has a transfer in space (rather than time) as can be seen in spatial validation (median KGE of 0.76 compared to median KGE of 0.97 in calibration; Fig. 3B vs. Fig. 3A).

Essentially, all these questions/suggestions are good but were not addressed and are out of scope for us. They might be worth a future inspection to see if the model performance could be improved by de-trending the forcings; a longer forcing dataset might be required though.

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?*

We totally agree with the reviewer that this is a very long manuscript. We were hoping that the clear structure and brief description of subsections at the beginning

of each section would be helpful for the reader to navigate through the manuscript and directly skip to the sections of interest. We tried to keep the sections as brief as possible without removing any detail that is required to follow the main analyses and conclusions. We do not want to move the (brief) model descriptions to the Supplements as this seems integral for a model intercomparison study. All details that are not regarding the following specifics have been already moved to the Supplements (the following list of bullet points was given to each collaborator contributing a section describing their model):

- **introductory sentence including major references for model and aim of model**
- **model resolution (spatial, temporal)**
- **used forcings and derived basin attributes**
- **calibration method (algorithm, objective, iterations/budget, independent trials, parameters go to Supplements)**
- **validation: donor basin mapping or regional/global setup?**
- **evaluation: model outputs AET, SSM, SWE (maybe add the actual model variable that is dumped to the output file; make note about SSM how this is modeled and why it was important to show only standardized SSM (i.e., only using correlation))**

We hope it is ok if we do not make any adjustments to the manuscript here.

References

Barbara Belletti, Carlos Garcia de Leaniz, Joshua Jones, Simone Bizzi, Luca Börger, Gilles Segura, Andrea Castelletti, Wouter van de Bund, Kim Aarestrup, James Barry, Kamila Belka, Arjan Berkhuysen, Kim Birnie-Gauvin, Martina Bussettini, Mauro Carolli, Sofia Consuegra, Eduardo Dopico, Tim Feierfeil, Sara Fernández, Pao Fernandez Garrido, Eva Garcia-Vazquez, Sara Garrido, Guillermo Giannico, Peter Gough, Niels Jepsen, Peter E Jones, Paul Kemp, Jim Kerr, James King, Małgorzata Łapińska, Gloria Lázaro, Martyn C Lucas, Lucio Marcello, Patrick Martin, Phillip McGinnity, Jesse O'Hanley, Rosa Olivo del Amo, Piotr Parasiewicz, Martin Pusch, Gonzalo Rincon, Cesar Rodriguez, Joshua Royte, Claus Till Schneider, Jeroen S Tummers, Sergio Vallesi, Andrew Vowles, Eric Verspoor, Herman Wanningen, Karl M Wantzen, Laura Wildman, and Maciej Zalewski. More than one million barriers fragment Europe's rivers. Nature, 588 (7838):436–441, December 2020.

US Army Corps of Engineers: Detroit District. Great Lakes Update – Volume 206: From Record-Lows to Record-Highs in 6 years. `https://www.lre.usace.army.mil/Portals/69/docs/GreatLakesInfo/docs/UpdateArticles/update206.pdf?ver=2020-07-01-115844-313`, 2020. Accessed: 2022-06-01.