Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/hess-2022-276-RC3, 2022

Comment on hess-2022-276

Anonymous Referee #3

Referee comment on "Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States" by Erin Towler et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-276-RC3, 2022

This is a review of the manuscript "Benchmarking High-Resolution, Hydrologic Performance of Long-Term Retrospectives in the United States" by Towler et al. The manuscript compares the performance of two large-scale hydrologic models in estimating streamflow by comparing against observed streamflow at gauges across continental United States (CONUS). The performance is evaluated using a number a metrics that are commonly used in streamflow evaluation. The manuscript is well-written and easy to follow. The effort to create benchmarks for CONUS scale streamflow prediction models is commendable, necessary, and of interest to this journal and the hydrologic community. However the metrics presented here are commonplace and the evaluation/benchmarking workflow is not novel. My biggest criticisms of the study are regarding the consistency of comparing two model outputs (major comment 2) and the use of calibration gauges in evaluation (major comment 3).

The manuscript can still be considered for publication provided the authors sufficiently address my concerns. I, therefore, recommend Major Revision.

Major comments:

- Introduction: The Introduction is missing a comprehensive review of current literature and needs improvement to further clarify the hurdles being overcome by this study and bring out its novelty. Specifically, the last paragraph should have a few sentences summarizing how it is building on previous studies and what shortcomings are being overcome in this specific study. Additionally, for studies mentioned in L 48-65, please mention their drawbacks and how this study aims to overcome them. Also, review of studies regarding statistical design of large-sample benchmarks and intercomparisons has been ignored. The authors should also clarify how the benchmark statistical design

used in this study compares to previous studies where large sample intercomparison and/or benchmarking have been carried out. Finally, the National Hydrologic Model is mentioned for the first time in the manuscript in L 75 when the authors are specifying the objectives of the study. The authors should introduce the two models briefly in the Introduction while also mentioning the reasons behind choosing these two specific models.

- L 113: NWM produces hourly streamflow using hourly atmospheric forcings whereas NHM produces daily streamflow using daily forcings. The hydrologic processes in the watersheds are simulated at different temporal scales (hourly vs daily) by the two models. Additionally, the many USGS gauges record 15-minute streamflow data. NWM can produce hourly streamflow and takes into account changes in hydrologic variables throughout the day. Averaging out higher resolution (hourly) streamflow timeseries produced using higher resolution (hourly) forcing to a coarser resolution is not the equivalent of simulating streamflow at a coarser resolution (daily) from coarse resolution (daily) forcings due to the non-linear nature of hydrologic processes. As such, is the comparison of the streamflow produced at two different temporal scales a consistent and fair comparison?
- Calibration: What was the calibration period for the two models? It is unclear from the text if gauges used in calibration were also part of evaluation. If the calibration period overlapped the evaluation period (October 1, 1983, to December 31, 2016), then the gauges used for calibrating either of the models should be removed from the set of gauges used for benchmarking the models. Including these gauges will introduce biases in the evaluation process.
- The study also includes gauges near the coast in the evaluation scheme. USGS gauges do not measure streamflow directly, rather the water surface elevation (WSE) is measured and the WSE is converted to streamflow using rating curves. Gauges near the coast can experience backwater from coastal surge traveling up the river and/or tides. In such cases, the rating curve for converting WSE to streamflow are violated and streamflow readings are highly erroneous. As such, should gauges near the coast be included in the evaluation scheme? Additionally, both NWM and NHM do not take into account the interaction between the river and sea/ocean.
- L 327-330: The authors should discuss why these areas are exhibiting poorer/better performance for both the models. They have a done good job of explaining the behavior of PBIAS in L 335-348 and need to similarly delve deeper into the potential causes of the behavior in the efficiency metrics for these regions.
- The authors need to discuss the limitations of this study and future work at the end of the manuscript in more detail. The limitations of the study extend beyond the subjectivity in choosing the performance metrics and their sensitivities. This could be a separated section or can be a continuation of the Results and Discussions.

Minor Comments:

- Title: is it really the United States if Alaska and the US territories have not been included? Should it be CONUS instead?
- L 177: The study uses 5,390 gauges and 5,389 of those are in GAGES II. So, there is just one gauge that was not part of GAGES II?
- L 191: "For statistical significance …" – statistical significance of what?
- L 350: refer to the appropriate table/figure
- Table 3 can be moved to supplementary information. KGE and NSE (and logNSE) are expected to behave somewhat similarly given their formulations. So this table does not convey anything particularly novel or important.

- Figure 2: There can be further subplots showing the CDF of KGE for the two models by region. This will be more informative than Table 4 which can then be moved to supplementary information.
- Figure 4: Just a suggestion, with there being so many points, it is hard to discern a trend or behavior from the figure. It might help to have region-wise or HUC-unit-wise medians color coded across CONUS. See Figure 8 in https://doi.org/10.1016/j.jhydrol.2022.127470 as an example.
- Please adjust the font size in the figures to make sure the legends, subplot number and lat/long are easily readable (Figures 3, 8, 11)