RC3: 'Comment on hess-2022-276', Anonymous Referee #3, 03 Oct 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.

General Reponse to Referee #3: We thank the Referee for these comments. We note that we provide a General Response for All Referees, which should be read first. We also respond here to each individual comment in this point-by-point response. As noted in the General Response, we have made major revisions to the (i) Introduction, (ii) Evaluation Approach (iii) Discussion and (iv) Results.

Major comments:

1. 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.

Thank you for these comments. We have revised the Introduction, please see the General Response. In short, we note that a drawback of most studies to date is that they are evaluating smaller, minimally-impacted basins (and we have added additional studies here, including Duan et al. 2006 using MOPEX, and Knoben et al. (2020) using CAMELS). Nevertheless, most river basins are impacted by human activities. These impacted basins also need to be benchmarked; especially as model development moves to include human systems. While there are some studies that have begun to address this globally (Arheimer et al. 2020); in Great Britain (Lane et al. 2019); Great Lakes Region (Mai et al. 2022); and for 1 year over the Central US (Tijerina et al. (2021); this has not been done for a long-term retrospective over the entire CONUS. This comprehensive evaluation of a long-term retrospective over the CONUS, using both impacted in addition to non-impacted sites, is our first contribution. We now note in the updated draft Discussion that to our knowledge, this is the first time that these models have been evaluated so comprehensively, Further, our second, related, contribution is facilitated by adopting and extending another suggestion to our paper, which was to provide performance context for our models. We now compare our two models to a climatological benchmark of KGE based on the interannual mean for each site, as in Knoben et al. (2020), and extend this by using it as a threshold to further scrutinize the metric results. Please see the General Response for details.

2. 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?

Thanks you for this comment. We note that in our original preprint, we did focus on the comparison between the NHM and NWM, whereas in the updated manuscript, we now compare both models with a climatological benchmark. However, we agree that we need to be transparent about the model differences, including the different temporal scales. This comes out in a new point brought up in the draft updated Discussion, where we now note: Another interesting difference in PBIAS was seen in the Central US, where the NHMv1.0 is underestimating volumes at underperforming sites. As detailed in the model descriptions, the model applications are run at different temporal scales: NHMv1.0 is run daily, whereas NWMv2.1 is run hourly and aggregated to daily. One hypothesis is that some precipitation events that are occurring on sub-daily scales, like convective storms, may be missed, or the associated runoff modes (Buchanan et al. 2018; https://doi.org/10.1002/hyp.13296). Similarly, while both models tend to underestimate high flows (PBIAS_HF) and variability (rSD), this is more pronounced for the NHMv1.0, which is in line with this hypothesis. The model applications showed differences in PBIAS_LF, with the NWMv2.1 overestimating low flows, whereas while the NHMv1.0 both over- and under-estimated them it was less extreme. It can be noted that both models used in the applications benchmarked here have only rudimentary representation of groundwater processes. Additional attributes (e.g., baseflow or aridity indices) could be strategically identified to further understand these model errors and differences. Model target applications, which drive model developer selections for process representation, space and time discretization, and calibration objectives, also have a notable imprint on the performance benchmarks. The NWMv2.1, with a focus on flood prediction and fast (hourly) timescales, shows better performance in high-flow-focused metrics, while the NHMv1.0, designed for water availability assessment and slower (daily) timescales, shows better performance in low-flow-focused metrics.

Buchanan, B., Auerbach, D.A., Knighton, J., Evensen, D., Fuka, D.R., Easton, Z. Wieczorek, M., Archibald, J.A., McWilliams, B., and Walter, T.: Estimating dominant runoff modes across the conterminous United States, Hydrological Processes, 32: 3881–3890, https://doi.org/10.1002/hyp.13296, 2018.

3. 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 calibration periods differed for the two models. For the NWMv2.1, the calibration period was from water years 2008 – 2013, and 2014-2016 was used for validation. For the NHMv1.0, the calibration period included the odd water years from 1981-2010, and the even water years from 1982-2010 were used for validation. This has been added to the model descriptions. As such, the calibration period did overlap with the evaluation period for the calibration gages, but it was not consistent between the models. We acknowledge the reviewer's point, but note that our approach fit with our objectives, which was to evaluate the long-term performance of both models at the same sites and time periods. The same technique was adopted in the MOPEX study (Duan et al. 2006). Further, there has been recent research activity in calibration. In particular recent studies suggest updating calibration techniques to use the full available data period and to skip model validation entirely (Shen et al. 2022; https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031523). We have added sentences to this effect in the updated Discussion section.

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. Water Resources Research, 58, e2021WR031523.

4. 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.

We do not include tidally influenced gages in this analysis, though this was an interesting point we had not considered. For the benefit of the reviewer, we note that we did speak to the USGS team, and they indicated that backwater effects are accounted for in USGS gauging procedures (tidal or otherwise). We believe that most are not stage-discharge gages but rather index-velocity gages which allow negative velocities. They indicated that the tougher issue is how to handle negative velocities (flows) in a hydrologic (water only downhill) type model; which would be challenging to either the NHM or NWM. In some previous work, when all the GAGESII references gages were analyzed, negative flows were very rare at the daily time step; recalling there were only two gages in Florida that had such. In short, we don't believe this to be a major issue in this analysis.

5. 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.

Thank you for these comments. We have updated our Results significantly to delve deeper into the performance of both models as compared to the climatological benchmark; please refer to the General Response, Results section.

6. 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.

We have updated the Discussion, including points raised in major comments #2 and #3. Further, we have added this as our updated Discussion final paragraph: "In closing, this paper uses the climatological seasonal benchmark as a threshold to screen sites for each model application. While this fit with the purpose of this study, the metrics for NWMv2.1 (Towler et al. 2022a) and NHMv1.0 (Towler et al. 2022b) are available for all sites (Foks et al. 2022); these can be analyzed and/or screened as needed. In the future, it would also be useful to extend the analysis beyond streamflow to other water budget components to assess additional aspects of model performance."

Minor Comments:

7. Title: is it really the United States if Alaska and the US territories have not been included? Should it be CONUS instead?

We have amended the title to be contiguous United States

8. 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?

Yes, only one was not part of GAGES II, but it fit all the other criteria so it was included.

9. L 191: "For statistical significance …" – statistical significance of what?

We agree that in the first draft of the paper this was confusing and unnecessary to the evaluation analysis. In updating our paper to compare both models with the climatological benchmark, we have removed this in the manuscript, see General Response.

10. L 350: refer to the appropriate table/figure

This has been fixed.

11. 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.

We have removed Table 3, and further now focus the paper on KGE (we have removed NSE and logNSE); see General Response.

12. 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.

Thank you for this comment – we agree and have changed several of the Tables to be CDFs, which we agree are much clearer; see the Results in the General Response.

13. 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.

Thank you for this suggestion. We note that we have removed Figure 4, and are now filtering the results by sites that underperforming relative to the climatological KGE benchmark (which reduces the number of points).

14. 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)

Some of the figures have changed, but we have adjusted the font sizes in the updated figures to make them easily readable. Further, we hope this will be helped by now stacking the figures vertically (rather than side-by-side).