RC2: 'Comment on hess-2022-276', Robert Chlumsky, 29 Sep 2022

I have completed my review of the paper "Benchmarking High-Resolution, Hydrologic Performance of Long-Term Retrospectives in the United States", Erin Towler et al. The paper presents a benchmark statistical design for the evaluation of process-based hydrologic models over large spatial and temporal scales, and is applied to evaluate the National Water Model v2.1 application of WRF-Hydro and the National Hydrological Model v1.0 of the Precipitation-Runoff Modeling System.

The paper itself is relatively straightforward in methods and application, including a description of both models, description of the metrics selected for evaluation and the presented comparison of the two models using the metrics selected. The paper draws a number of appropriate conclusions regarding the relative performance of the models spatially and based on flow regime, and is overall very well written and logically presented.

Regarding the comments on paper type, the paper aligns largely with a Technical Report format, though the additional discussion and interpretation of results help move it towards a Research Article.

A number of additional comments and concerns are presented here to help improve the paper.

General Reponse to Referee #2: 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; this has made our study now of more general interest, and suited to be a HESS Research Article; we appreciate the Referee comments towards this end.

General Comments

1. In the Introduction, mention of previous studies that have addressed the 5,390 USGS gages used in this study would be relevant (have any studies used all of these gages as well?)

To our knowledge, this is the first time these 5390 gages have been used in this type of comprehensive evaluation. We point out that we have substantially revised the Introduction, please see General Response.

2. Introduction - It would also be worth mentioning other datasets that have been commonly used in larger-scale benchmark and model intercomparison studies, such as the MOPEX (Duan et al., 2006) and CAMELS dataset (Addor et al., 2017). The Mai et al. (2022) GRIP-GL comparison would also be worth mentioning in the list of recent benchmark and model intercomparison studies.

We appreciate these suggested references and have added them to our revision. Please see General Response for the changes to the Introduction.

3. Introduction - Any previous studies benchmarking these two hydrologic models would be worth mentioning in the last introductory paragraph (lines 73-81), or mention that this is the first study benchmarking these two models specifically.

Although there have been some internal evaluation efforts, they have not been published. We now mention that this is the first time publishing the benchmark results for these two models, in the Discussion: "The presented analysis documented model performance for two large-sample, high-resolution hydrologic models over a long-term period. To our knowledge, this is the first time that these models have been evaluated so comprehensively, as this analysis included 5390 gages, both impacted and non-impacted by human activities."

4. Line 210 – are these three metrics providing very similar information for overall performance assessment in general, or simply because these models are similar and that happens to be the case in this study only? I would be surprised if this conclusion was generalized for very different hydrologic models, and I think this should be carefully rephrased to not overgeneralize from the limited model comparison (i.e. 2 similar models) presented in this study.

As indicated in our General Response, we removed the NSE and logNSE metrics from the suite. As such, we have removed the lines in question from the manuscript.

5. Reference to Knoben et al. (2019) on what a baseline KGE performance is may be useful in interpreting the results, since 0.2 seems somewhat artbitrary. The Knoben et al. paper suggests -0.4 is a more comparable threshold to the NSE=0 interpretation, so perhaps some justification or rationale for using 0.2 is warranted.

We agree with this comment, and as indicated in our General Response, we are now use the KGE > –0.41 as the mean flow benchmark (Knoben et al. 2019), as well as computing the interannual mean/median benchmark values as in Knoben et al. (2020).

6. Table 4 – the bolding pattern is confusing to me, since it is meant to represent the maximum number (percent) of sites by KGE category (?), though the Northeast has two bolded numbers, and in the Central region the minimum number is bolded. Similar bolding patterns continue in other Tables and seem to be at least non-intuitive.

Thanks for this feedback. In our revision, we have removed most of the tables, and have replaced them with CDFs that better show the differences between models, regions, and classification. See the Results in the General Response.

7. Table 6 – I would suggest a summary column with the average metric across regions to help summarize the results, similar to how Table 5 summarizes results for Ref and non-

ref sites. This would have some duplication with Table 5 but I think it is still worth including here as an additional column

Thank you for pointing this out. Similar to the previous comment, we have removed this table in light of our updated manuscript. Please see updated Results in the General Response.

8. Figure 4 and lines 241-247 – I thnk that screening the models with poor initial performance from Figure 4, perhaps as a separate figure, would be more meaningful than comparing relative model performance between a KGE of 0.0 and -0.05. In either case, the models likely don't capture enough of the observed behaviour for a modeler to care which is better, and this inhibits interpretation of Figure 4 in identifying any real differences between the models. It seems the models will be similar in any case, but I would filter results first.

Agreed, we have taken this suggestion to heart and adopt this approach of screening the models with poor performance relative to the KGE benchmark. Please see General Response.

9. Line 268 – I think this statement is actually incorrect, since the lower variability at managed (non-reference) sites should already be normalized by comparison to observed data. My interpretation of this is that the ideal rSD is 1.0, and rSD below 1.0 indicates that the model underestimates the variability of flow. In both cases the models underestimate the variability of flow, in particular for reference sites relative to non-reference or managed sites. This suggests the models do better at capturing general changes in flow rather than sudden ones in unregulated reference sites perhaps. There is more interesting interpretation to add in this section.

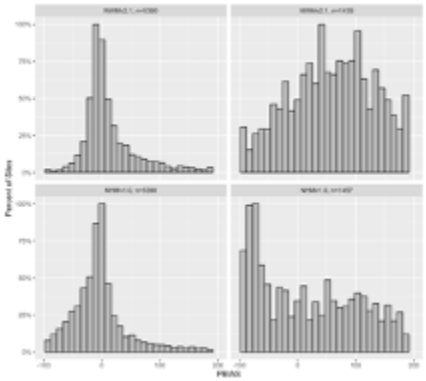
We have revised the Results (see General Response), and have revised our examination of the rSD results.

10. Line 277-279 – this can be compared with the GRIP-GL study results (Mai et al., 2022) to discuss general trends in Great-Lakes areas

Thanks for this suggestion. We appreciate pointing out the Mai et al., 2022 paper, and have added it to our Introduction, specifically where we discuss studies that include both impacted and non-impacted gages (See General Response). Further, given that we revised our study to compare the NHM and NWM CONUS-wide models with climatological seasonal benchmark, it would have broadened the scope too much to further compare directly with outputs from other model studies.

11. Line 295 – general comment but an actual histogram plot of the information in Supplemental Table 2 would likely convey this information much better and would aid the discussion. A simple histogram of frequency vs binned PBIAS_LF, and either facet or colour code each of the four regions on one plot would greatly aid the discussion

Agreed, we have removed Supplemental Table 2 and have revised our results to better convey the information, both in terms of new CDF plots and by filtering our results based on the climatological benchmark. We have also added histograms of some of the metrics (PBIAS, PBIAS_LF, PBIAS_HF, and rSD) for both models as separate figures in the Supplemental Material, which are referenced in the updated Results (See General Response). For the benefit of the reviewer, we include the histogram of PBIAS as a preview of what will be included in the finalized Supplemental Material:



Supplemental Figure 4: Normalized histograms of PBIAS for National Water Model v2.1 (NWMv2.1, top) and National Hydrologic Model v1.0 (NHMv1.0, bottom), for all sites (left) and for sites where the model's KGE score is less than the average day-of-year flow benchmark (right).

12. Line 341 – it would be worth elaborating on the value of the passive lake/reservoir representation in the model relative the none. It is interesting that the model with the passive representation (NWMv2.1) does seem to perform slightly better than the NHMv1.0, though it is unclear if that is the reason why or what the improvement in performance would be with a better representation of reservoir operations. This would require some segmentation based on catchments with 'significant' reservoir controls, which is not included in this study, though worth discussing briefly here.

Agreed, these differences in performance were only slight, and in our revision, we focus less on the differences between the NHM and NWM. Given the overall manuscript changes, we don't add on to this point in the manuscript, but have added to the draft Discussion: "As model development moves towards including human systems, the benchmark results could potentially provide a more concrete goal for "how much" improvement would be needed to adopt a management module. This is of increasing interest as the hydrologic modeling community grapples with how to account for the anthropogenic influence on watersheds, especially since most studies to date focus on minimally disturbed sites."

13. Line 355 – the NWMv2.1 is described to perform better in high-flow-focused metrics than the NHMv1.0. This discussion should be expanded to how this could likely have

been known from the model setup initially, since running the model on hourly or subdaily timesteps and aggregating will very likely produce better performance for peak flow metrics than a model that is run at a daily timestep, therefore this result should not be a surprise. This is touched on by mentioning that the latter model is designed for water availability, but I think this point should be emphasized.

We have expanded on this, especially in light of new results showing the underestimation of PBIAS in Central by the NHM, we now note in the updated draft Discussion: "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). 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."

14. Conclusion – the concluding paragraph ends rather abruptly, a short one or two line paragraph at the end to tie off the accomplishments of the paper and goals for future studies would help to transition the conclusion.

Thanks for this suggestion. 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."

Technical Comments:

15. I was under the impression that CONUS was an acronym for contiguous United States (not conterminuous), though I suppose the definitions are practically the same

Thank you, this has been changed to contiguous.

16. Links in lines 92-93 should be properly cited instead of providing raw urls

This citation has been updated.

17. Line 168 – I would rewrite this paragraph slightly to something like: "Three additional hydrologic signatures are included which evaluate performance based on different parts of the flow duration curve (FDC) for high, medium, and low flows. The definitions for these hydrologic signatures as used in this study are consistent with those from Yilmaz et al. (2008). The bias of high flows…" This will help the readability of the section, otherwise the reader is left wondering which metrics you are porting in from Yilmaz until the whole section is read.

Thank you, this has been edited as suggested. We note that we now only include PBIAS_LF and PBIAS_HF (we have removed PBIAS_FDC).

18. Line 201 – "…for all 5,390 cobalt gages …". If these will be called cobalt gages in the paper, this should be used throughout the paper after its definition for consistency

We have removed the term cobalt gages throughout the manuscript.

19. Line 221 – the line "Both models also have many sites with poor performance" – this can be quantified and merged with the next line, as many sites in a large sample study could mean 100 or 1000. Both models in fact have 30% of their sites with a KGE below 0.2, which is a lot of models with very poor performance (KGE below 0.2 is likely an 'unusable' or 'untrustworthy' model for most applications)

These lines have been updated now that our analysis includes a comparison to the climatological KGE benchmark. See General Response.

20. Line 361 – link should be properly cited

In revising our Discussion, this has been removed.

References

Thank you for these references, they have been added.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology, 320(1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Knoben, Wouter & Freer, Jim & Woods, Ross. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrology and Earth System Sciences Discussions. 1-7. 10.5194/hess-2019-327.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W. (2022):The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL) Hydrol. Earth Syst. Sci., 26, 3537–3572. Highlight paper. Accepted Jun 10, 2022.