Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-276-RC1, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on hess-2022-276

Anonymous Referee #1

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-RC1, 2022

Review of "Benchmarking High-Resolution, Hydrologic Performance of Long-Term Retrospectives in the United States" by Towler et al.

Summary

In this paper, performance of the National Water Model (NWM) v2.1 and the National Hydrologc Model (NHM) v1.0 is evaluated over the United States. These models are different in their internal structure, use different calibration approaches and are run with different meteorological inputs, but are similar in the sense that both are run over a high-resolution spatial grid. Model performance is evaluated with the help of 9 different metrics (e.g. Nash-Sutcliffe, PBIAS) that are calculated using observations and model simulations at 5390 streamflow gauges. Attention focuses most on median values in the 5390-member sample, and on differences between both models in various broad regions across the US. There are some recommendations on how to improve both models; most notably by updating the model structures to account for human water use and the impact of lakes and reservoirs.

General comments

Having read this paper, I must admit that I am not entirely sure whether HESS is the right venue for this. Various sentences suggest that this publication is intended as a benchmark for further development of the NWM and HWM. For example:

- [line 25] "This benchmark provides a baseline to document performance and measure

the evolution of each model application"
- [line 80] "This paper highlights select results of the benchmarking analysis to document baseline model performance and characterizes overall performance patterns of both models."
- [line 198] "Here, we provide select results, with a focus on documenting baseline model performance and providing insight towards model diagnostics and development."
- [line 315] "here we provide a lower benchmark to gauge the evolution of the NWMv2.1 and NHMv1.0".

This is a great goal that I think should be the standard in any model development exercise (as it is in many other fields), but this kind of benchmarking is of limited interest to anyone who does not actively work with these models. A technical report instead of a journal publication might be more appropriate.

To appeal to a wider (international) journal audience, the proposed benchmarking approach should be of general interest and I think in its current shape it fails to be that. My main concerns are that:

- The selected benchmarking metrics are too one-sided: out of the 9 metrics, 7 either include or are some form of model bias metric. Multiple other relevant aspects of hydrographs and model performance are not captured by these metrics.
- There is no clear way to relate a model's performance on this set of metrics to concrete suggestions for improvement of the model, because it is practically impossible to trace the scores a model obtains on these metrics to how well the model simulates a given hydrological process (though I appreciate that this is not an easy thing to do).
- The model results are presented in a vacuum: there is only very limited discussion of existing literature on benchmarking, there is no comparison of the performance of these two models to the performance of earlier modeling efforts across this domain, and there is no discussion about how high a model must score on any of the 9 metrics to be considered a good/plausible/acceptable/etc model.
- There is almost no guidance (or better yet, software) available for a reader who might want to apply this benchmarking approach to their own simulations, beyond a table that shows references for the 9 metrics and a CSV file that contains the list of gauge IDs.

I believe that these issues can be addressed to a certain extent (see specific comments below), but in its current shape this manuscript mostly describes what performance scores two arbitrary models obtain on a limited selection of model performance metrics, without any context for those scores whatsoever. I don't think that's enough to warrant publication in HESS.

Specific comments

l12. "a benchmark statistical design" - It's unclear to me what this means.

l90. "https://noaa-nwm-retrospective-2-1-pds.s3.amazonaws.com/index.html" - The NWM docs (https://water.noaa.gov/about/output_file_contents) seem to say that output files are in netCDF4 format, but if I follow this link all I can find is .comp files. What are these files and how can a reader open/use them?

l105. "Using the AORC meteorological forcings, NWMv2.1 calibrates a subset of 14 soil, vegetation, and baseflow parameters to streamflow in 1,378 gauged, predominantly natural flow basins. The calibration procedure uses the Dynamically Dimensioned Search algorithm (Tolson and Shoemaker, 2007) to optimize parameters to a weighted Nash-Sutcliffe efficiency (NSE) of hourly streamflow (mean of the standard NSE and log-transformed NSE). Calibration runs separately for each calibration basin, then a hydrologic similarity strategy is used to regionalize parameters to the remaining basins within the model domain." - This needs a reference to indicate where a reader can find further details about this procedure.

l113. "For the analysis in this work, hourly streamflow is aggregated to daily averages." - Looking at a snapshot of the USGS gauges used for this evaluation approach, observations seem to be available at a sub-daily resolution. Given that the model is run at a 3-hr resolution, and it is known that hydrologic processes of interest can show strong diurnal variation (e.g. evaporation, snowmelt), why are observations and simulations aggregated to daily values?

l148. "The NSE is formulated to emphasize high flows" - This statement seems to contradict the last part of this sentence: "models do not necessarily perform well at reproducing high flows when NSE is used for calibration". Suggest to rephrase this.

l156. "Correlation, standard deviation ratio, and percent bias" - These three are (almost) the constitutive components of the KGE metric, and also appear in the NSE (see e.g. the decomposition of RMSE by Murphy, 1988, https://doi.org/10.1175/1520-0493(1988)116%3C2417:SSBOTM%3E2.0.CO;2). There is likely value in looking at these individual components compared to the aggregated efficiency scores, but this section should state that these metrics are not independent from NSE and KGE.

l167. "Three hydrologic signatures defined by Yilmaz et al. (2008)" - There are many possible signatures one could chose from and these are sometimes divided into five separate categories (magnitude, frequency, duration, timing and rate of change; e.g. Olden & Poff, 2003, dx.doi.org/10.1002/rra.770). More recently, McMillan (2022; dx.doi.org/10.1002/hyp.14537) created a signature taxonomy that relates signatures to specific hydrologic processes. The selected signatures here exclusively address the magnitude component, without explaining why these other components are not addressed or how a model's performance on any of these signatures might inform which of the

model's process representations needs to be improved.

More generally, out of the 9 presented metrics, 7 metrics are either some form of bias or include a bias component. This seems insufficient spread to me for a "standard metric suite". I believe this selection needs to be expanded quite a bit before these metrics can start to be used for comprehensive model benchmarking.

l170. "big precipitation" - This might be inaccurate phrasing in the case of colder catchments, where flow events might originate from snow/ice melt and not directly from individual precipitation events.

l178. "Foks et al., 2022" - The .csv file in this reference misses leading zeroes for station numbers, which makes searching for them somewhat difficult on the USGS website (https://waterdata.usgs.gov/nwis/uv?referred_module=sw&search_criteria=search_site_no&search_criteria=site_tp_cd&submitted_form=introduction). E.g. searching for station 1011000 yields no results with the default "exact match" option, whereas 01011000 does show a result. If possible, updating this resource could help others. Adding some guidance on how to obtain these observations in a reasonably efficient manner would be good too.

l191. "For statistical significance, we conduct pairwise testing, specifically the Wilcoxon signed-rank test. The Wilcoxon signed-rank test is a non-parametric alternative to paired t-test. The Wilcoxon signed-rank test is appropriate here since the metrics (particularly the efficiency metrics) contain outliers and are not necessarily normally distribute" - This is unclear to me. What is being compared pair-wise? Why? A reference to point the reader to info about a Wilcoxon signed-rank test would be good too.

l202. "median values" - Why are only medians discussed here? How meaningful is that on a 5000+ sample?

l206. "indicating that they are tracking similarly in terms of overall performance" - This may need to be a more nuanced. Because these correlations are calculated on ranks and not actual metric scores, I think all this indicates is that these models are similar in where they tend to do relatively better and worse (within their own 5390-member sample). I don't think these ranked correlations indicate that these models are similar in actual performance as measured by the metrics, which is what the text seems to say.

l209. "these three popular efficiency metrics are providing very similar information in terms of overall performance assessments" - Again, I think this may need to be a bit more nuanced. What I believe these correlations show is that relative ranks are similar for these three metrics. In the .csv files I can see that there are still quite large differences in the actual scores on the three metrics. I would suggest to rephrase this paragraph.

l216. "Figure 2" - Why is the x-axis in this figure capped at KGE = -0.25? Looking at the data in the .csv files I see that KGE scores go as low as KGE = -306 for the NWM, and KGE = -158 for the NHM. This suggests that there is a lot of rather poor model performance that's not shown in this figure. Should that not be discussed as well in a paper intended to set a baseline for model performance?

l219. "Table 4 bins the KGE scores" - A similar question can be asked here: why are these bins defined with a lower bin of KGE < 0.2? There seems to be a lot of variety in model performance below this arbitrary threshold. More generally speaking, what can be learned by binning the data in this way that is not obvious from a figure with four CDFs (one CDF each for west, central, southeast and northeast)? These KGE bin boundaries seem quite arbitrary to me and mask any variety within the bin. It might be cleaner to replace this table with CDFs per region instead.

l231. "Relatively good performance is seen in the Southeast" - This paragraph uses fairly arbitrary thresholds to discuss the KGE performance of both models (e.g., anything with KGE < 0.2 is considered poor performance; KGE > 0.8 is implicitly treated as a boundary above which everything is similarly good). Previous publications argue that efficiency scores such as NSE and KGE cannot be viewed in isolation but need to be compared to some form of baseline model, so that one can judge if these NSE/KGE scores are in fact poor or good for a given location (e.g. Seibert, 2001; Schaefli & Gupta, 2007; Pappenberger et al., 2015; Seibert et al., 2018). NSE includes such a benchmark by design (i.e. the mean annual flow - but this is often criticized as being too easy to beat). KGE does not include such a benchmark and therefore needs some other way to provide context. Work using the CAMELS catchments (Knoben et al., 2020) uses a seasonal cycle benchmark and suggests that for certain locations even KGE > 0.9 could be considered a basic requirement for models rather than being indicative of an exceptionally well-performing model. I think the KGE scores discussed in this paragraph need to be given some context, so that there is some objective reason to qualify a given KGE score as "poor", "good" etc. Presenting these scores in isolation does not help the reader understand what kind of model performance they indicate.

The same comment applies to the following paragraphs as well. The presented numbers need some context that gives the reader an objective reason to decide whether those numbers are indicative of good or bad model performance.

Knoben et al.: doi.org/10.1029/2019WR025975
Pappenberger et al.: doi.org/10.1016/j.jhydrol.2015.01.024
Schaefli & Gupta, 2007: doi.org/10.1002/hyp.6825
Seibert, 2001: doi.org/10.1002/hyp.446
Seibert et al.: doi.org/10.1002/hyp.11476

l244. "It is noticeable that many of the sites are in the tails" - I find this hard to grasp from just looking at this figure. Adding a small histogram to the bottom left corner might help.

l315. "here we provide a lower benchmark to gauge the evolution of the NWMv2.1 and NHMv1.0" - This sentence seems to suggest that this publication is mainly intended to benchmark future development of the NWM and NHM. Would a technical report not be a more appropriate venue for this? The kind of information presented in this paper seems useful to those actively working with the NWM or NHM, but may be of somewhat limited interest to the wider hydrological audience.

l317 "The baseline can provide an a priori expectation for what constitutes a “good” model." - I respectfully disagree. This baseline shows the current performance of the NWM and the NHM but it provides no objective reason for calling either a good model. For example, the mean annual flow (NSE = 0; KGE = -0.41) is often used as a rudimentary threshold for model performance. The .csv files with metric values show that the NWM does not outperform the mean annual flow as a predictor in 23% of gauges if NSE is used, and 14% of gauges if KGE is used. Similarly, the NHM does not outperform a mean annual flow in 24% of cases if NSE is used, and 12% of cases if KGE is used. To make the statement that these results are a priori expectations for what constitutes a good model, a much more in-depth comparison of both models against a range of statistical benchmarks (e.g., mean annual flow, seasonal cycle, persistence) and existing model results across this domain (e.g. any number of results based on the CAMELS data, NLDAS [10.1029/2011JD016051], global models [10.5194/hess-24-535-2020]) is needed.

l336. "Results helped to identify potentially missing processes that could improve model performance. PBIAS results showed that for both models, simulated streamflow volumes are overestimated in the West region, particularly for the sites designated as Non-Reference. One primary reason for this may be that water withdrawal for human use is endemic throughout the West and neither model has a thorough representation of these withdrawals. Furthermore, neither model possesses significant representations for lake and stream channel evaporation which, through the largely semi-arid west, can constitute a significant amount of water "loss" to the hydrologic system (Friedrich et al., 2018). Lastly, nearly all western rivers are also subject to some form of impoundment. Even neglecting evaporative, seepage and withdrawal losses from these water bodies, the storage and timed releases of water from managed reservoirs can significantly alter flow regimes from daily to seasonal timescales thereby degrading model performance statistics at gaged locations downstream of those reservoirs" - Upon reading this I cannot help but wonder if PBIAS values were needed at all to determine that these models might be improved by accounting for human water use and the presence of lakes & reservoirs. These seem fairly obvious processes to me when one is working with "two models that have been developed to assess water availability and risks in the United States". Should this even be listed as a discussion/conclusion point, instead of being presented as a known a-priori limitation of these models?

l357. "state-of-the-art" - Without intending to disparage the work that undoubtedly has already gone into creating these models, calling them state-of-the-art seems an overstatement if neither of these water resources assessment tools has a way to account for human interaction with the water cycle.

l354. "Identifying a suite of metrics has an element of subjectivity, but our aim was to identify an initial set of metrics that can be applied to a wide variety of science questions

(e.g., see Table 1.1 in Blöschl et al. 365 2013) and that build on standard practices for evaluation of model application performance within the hydrologic community" - As indicated earlier, with 7 out of 9 metrics focusing on bias I find this set of metrics too limited for even an initial set. Of course there is some subjectivity in selecting metrics, but there is also some existing understanding of which statistical properties of hydrographs might be relevant to look at, how those might be captured in streamflow signatures, and how those signatures might be used to explain how well a model simulates certain, specific processes. This current selection of metrics seems too ad-hoc to me and some deeper literature searching would likely result in a set of metrics with a much wider applicability.

l576. "Table 1" - It would be helpful if equations were added to each row here. The ratio metrics are currently difficult to interpret for the reader, because they cannot know whether these are calculated as sim/obs or obs/sim without looking into other references.

l576. "Table 1" - Why are these bias metrics capped at (-)100?

l642. "Reference (Ref, n= 1,115) and Non-Reference" - A brief explanation of what reference/non-reference means would be helpful. This could be a summary of lines 186-189).

Technical corrections

l162. "modeled and observed" - Is there a word missing that should come after "observed"?

l197. "Using daily observations and simulations from the NWMv2.1 (Towler et al., 2022a) and NHMv1.0 (Towler et al., 2022b)
hydrologic modeling applications" - The way the Towler et al. references are inserted in the text implies that they contain the daily time series of observations and simulations, but in reality these references include only the 9 metrics for each gauge. Suggest to clarify this.

l204. "the differences are statistically significant given the large sample size" - Why are some values bold in the NWM column and others in the NHM column? Shouldn't they be bold in both or neither?

l230. "you move" - consider replacing with "one moves"

l241. "better and worse" - is there some text missing here that indicates compared against what these models do better or worse?

l403. "References" - This list is not entirely in alphabetical order.

l557. "https://10.5066/P9DKA9KQ" - Has this link been inserted correctly? When I click it it attempts to take me to a local file location instead of the link the text suggests this is. Unsure if this problem is on my end only, but the link in the Towler reference above this one seems to work fine for me.

l644. "Figure 2" - these figures are quite small. Stacking the subplots vertically would give more space to each figure.

l673. "Figure 8" - these figures are quite small. Stacking the subplots vertically would give more space to each figure.

l687. "Figure 11" - these figures are quite small. Stacking the subplots vertically would give more space to each figure.