Referee #1.

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.

We disagree in that this shows how a benchmarking approach can be used for additional modeling applications.

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.

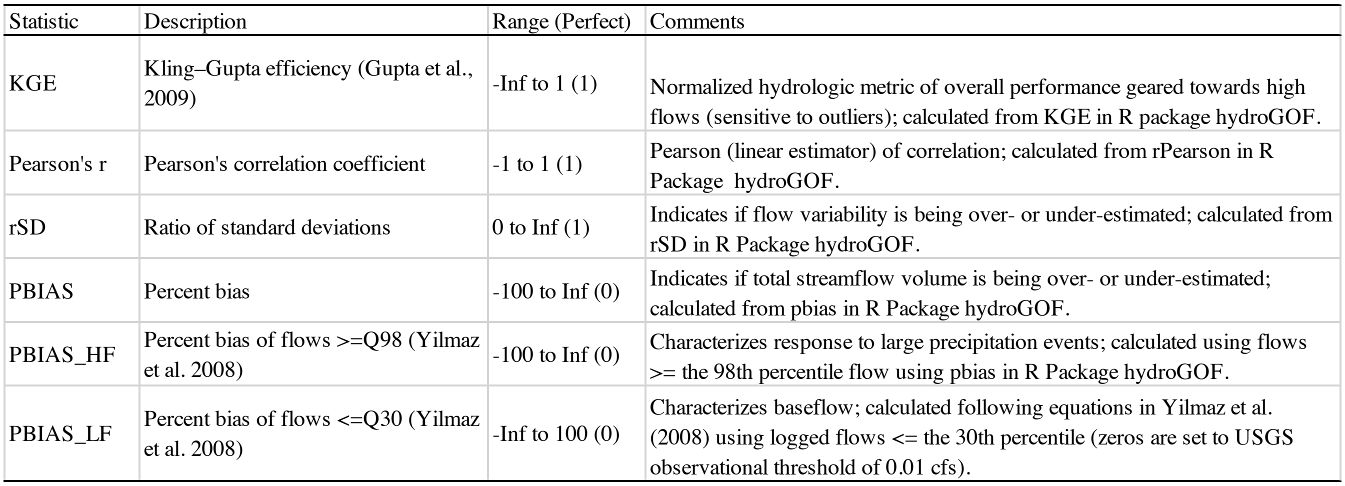
General Reponse to Referee #1: We thank the Referee for these comments. We note that we had an initial "Short Response" to your general comments, and now we have fleshed out our response in the "General Response for All Referees", which should be read first. We also respond here to each individual comment 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 a HESS Research Article, and we appreciate the Referee comments that helped us to achieve this end.

My main concerns are that:

1. 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.

Please see "Evaluation Approach" in the General Response. As indicated, we now focus on KGE and reduce the number of metrics examined (see Table 1 in General Response, and included here):

Table 1. Standard metric suite included in the daily streamflow evaluation. KGE = Kling–Gupta efficiency; Pearson's r = linear correlation; rSD = ratio of standard deviations between simulations and observed; PBIAS = percent bias; HF = high flows; LF = low flows.



Statistic	Description	Range (Perfect)	Comments
KGE	Kling–Gupta efficiency (Gupta et al., 2009)	-Inf to 1 (1)	Normalized hydrologic metric of overall performance geared towards high flows (sensitive to outliers); calculated from KGE in R package hydroGOF.
Pearson's r	Pearson's correlation coefficient	-1 to 1 (1)	Pearson (linear estimator) of correlation; calculated from rPearson in R Package hydroGOF.
rSD	Ratio of standard deviations	0 to Inf (1)	Indicates if flow variability is being over- or under-estimated; calculated from rSD in R Package hydroGOF.
PBIAS	Percent bias	-100 to Inf (0)	Indicates if total streamflow volume is being over- or under-estimated; calculated from pbias in R Package hydroGOF.
PBIAS_HF	Percent bias of flows >=Q98 (Yilmaz et al. 2008)	-100 to Inf (0)	Characterizes response to large precipitation events; calculated using flows >= the 98th percentile flow using pbias in R Package hydroGOF.
PBIAS_LF	Percent bias of flows <=Q30 (Yilmaz et al. 2008)	-Inf to 100 (0)	Characterizes baseflow; calculated following equations in Yilmaz et al. (2008) using logged flows <= the 30th percentile (zeros are set to USGS observational threshold of 0.01 cfs).

In the Discussion, we now also acknowledge that we focus on magnitude, since one of the main purposes of the model evaluation was to assess the suitability of the models for water availability studies. However, we now note that magnitude is only one aspect of streamflow, and that different metrics for other categories (e.g., frequency, duration, rate of change, etc), could be more appropriate for addressing specific modeling objectives.

2. 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).

We agree that this is a difficult endeavor. Nevertheless, we liked the suggestion of adopting the climatological benchmark of Knoben et al. 2020, since it offers a concrete goal for model development. Further, by screening our results to look mainly at the underperforming sites (I.e., sites that have KGE values below the climatological benchmark), we were able to come up with several hypothesis as to why, which could be useful for model development.

3. 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.

As seen in our General Response, we have done a major revision, and adopted the suggested approach of comparing with a climatological seasonal benchmark, and using this as a threshold to screen the results.

4. 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.

As mentioned in the initial Short Response, most of the metrics are straightforward to calculate (we use HydroGOF in R), and we have also added the Equations to the manuscript. Based on other comments below, we clarify the utility of the published metrics, and we update them by adding the climatological mean/median benchmark for each site.

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.

Please see General Response for the Major Revisions applied to address the Reviewer's comments.

Specific comments

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

We have removed this term from the text.

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?

The output files are in netCDF files, they are just tagged as ".comp" because they are compressed. We will add to the manuscript: "(e.g., compressed netcdf files can be found at:…". The netCDF package in R allows for opening and viewing of netcdf files, but a reader can use a variety of programs to open these files.

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.

At this time, there is no publication to reference on this, but authors on this publication provided additional details. Based on this and other Referee comments, we have added the calibration periods to the model descriptions. 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.

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?

Not all of the gages contain sub-daily records for the temporal extent of interest (1983-2016). Additionally, the NHM only can simulate streamflow at the daily timestep and comparison of these two models on different timesteps was not appropriate. For the benefit of the Referee, we note that other internal evaluations of the NWM have been conducted hourly, but that wasn't the focus for this study.

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.

In our revision, we have removed NSE (and logNSE) to focus on KGE results, and the new climatological benchmark comparison. As such, this sentence has been removed.

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.

Thank you for raising this point. In our revision, we are focusing on KGE, and agree that there is value in looking at its constituents. We now look at the components of linear correlation (as opposed to in the previous draft, we looked at Spearman's correlation), standard deviation ratio, and percent bias. These equations are now included in the manuscript.

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.

We appreciate this comment and these references, this was part of the impetus towards our Major Revision (see General Response). We have added draft material to the Discussion to address this point explicitly in the paper: "Identifying a suite of evaluation metrics has an element of subjectivity, but our aim was to focus on streamflow magnitude, since the purpose of the model evaluation effort was for water availability applications. However, magnitude is only one aspect of streamflow, and different metrics for other categories (e.g., frequency, duration, rate of change, etc) could be more appropriate for addressing specific scientific questions or modeling objectives. Recently, McMillan (2019) links hydrologic signatures to specific processes using only streamflow and precipitation. Interestingly, McMillan (2019) does not find many signatures that relate to human alteration; however, in this paper, streamflow bias metrics are found to be useful in this regard."

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.

Thank you for this comment. We have added "big precipitation or melt 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.

We are not sure what software you used to open the CSV files, but a text-editor such as Rstudio, Notepad++, Visual Studio are common to use for opening CSV file formats so that leading zeros are observable. Microsoft Excel truncates data types it assumes are numeric values. The metadata accompanying this release has information regarding the leading zero for the station IDs.

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.

We agree that the statistical significance analysis we included was not clear in the first draft, and not altogether necessary for the paper. In updating our paper to compare both models with the climatological benchmark, we have removed this (formerly Table 2) in the manuscript. See General Response Results section for more information.

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

In our updated analysis, although we still sometimes provide the median for a quantitative point-of-reference, we now include CDFs (Cumulative Density Functions) of the KGE results for (i) the models and the climatological benchmarks, (ii) the Reference versus Non-Reference classification of the gages, and (iii) the 4 regions. See the figures in the General Response Results section for more information.

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.

Thank you for this comment. In our revision, we have replaced the Spearman rank correlation with the linear correlation, so to be more consistent with the components of KGE. As such, we have removed this sentence.

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.

Thank you for this comment. In our revision, we have removed the NSE and logNSE from the manuscript, so as to focus more on KGE and its comparison with the climatological benchmark. As such, we have removed this sentence.

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?

This is a good point, and as indicated in our General Response, we are now comparing with climatological benchmarks, including the mean annual flow KGE benchmark, i.e., –0.41 (Knoben et al. 2019). We have adjusted our x-axis to include this in Figure 2, 3, and 4 (see General Response).

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.

Thank you for this point and suggestion. We have added a new figure that shows the CDFs by region (Figure 4) and removed the previous Table 4 (see General Response).

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

We appreciate the reviewer's comments here, as well as the literature suggestions. As seen in the general response, we have taken this comment to heart, and are now comparing with the climatological seasonal benchmark following Knoben et al. (2020). We have also added the suggested literature to provide more background for our work.

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.

In our revision, this sentence has been removed, as we have removed this Figure (formerly Figure 4) and now show the difference between the in Kling–Gupta efficiency (KGE) from the maximum model (i.e., the maximum from the NWMv2.1 or the NHMv1.0) minus the seasonal benchmark based on the average day-of-year flows. See General Response Results for more information.

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.

Thank you for this comment, as we noted in our General response, we appreciate the suggestions to compare with a climatological KGE benchmark to make this of greater interest to the wider community. See General Response for more information.

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.

Thank you for this comment. We have revised our paper to compare both models against the climatological KGE benchmarks, including mean annual flow and mean/median daily of year flows). We also appreciate these additional references, and have added them to our Introduction (see General Response).

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?

In our revision, we have added more in our Discussion around this point. We note that 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. It is also interesting to see that PBIAS is the component that is most useful for this aspect of model diagnostics. Recently, McMillan (2019) links hydrologic signatures to specific processes using only streamflow and precipitation. Interestingly, McMillan (2019) does not find many signatures that relate to human alteration; however, in this paper, streamflow bias metrics are found to be useful in this regard.

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.

This has been removed.

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.

Thank you for this suggestion; in addition to our General Response, we provide this draft excerpt to be added to the Discussion: “Identifying a suite of evaluation metrics has an element of subjectivity, but our aim was to focus on streamflow magnitude, since the purpose of the model evaluation effort was for water availability applications. However, magnitude is only one aspect of streamflow, and different metrics for other categories (e.g., frequency, duration, rate of change, etc) could be more appropriate for addressing specific scientific questions or modeling objectives. Recently, McMillan (2019) links hydrologic signatures to specific processes using only streamflow and precipitation. Interestingly, McMillan (2019) does not find many signatures that relate to human alteration; however, in this paper, streamflow bias metrics are found to be useful in this regard.”

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.

We have added the equations to the manuscript, and see next response.

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

This is the range for PBIAS as we define it; we have added the equation to the paper to make this clear (and equations for PBIAS_HF and PBIAS_LF to the Supplemental):

Percent bias (PBIAS) is calculated as (Zambrano-Bigiarini 2020):

$$PBIAS = \frac{\sum_{t=1}^{N}(S_t - O_t)}{\sum_{t=1}^{N} O_t}$$

where observed flow is *O*, simulated flow is S, and t = 1, 2,… N is the time series flow index.

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).

Ref and Non-Ref are now defined in section 3.1. Data, and used consistently throughout.

Technical corrections

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

We have added “streamflow” after.

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.

This has been clarified.

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?

We have removed the statistical significance analysis from the paper.

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

This has been replaced.

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

This sentence has been removed in light of our Major Revision.

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

Thank you, we will check the references for 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.

This has been updated and should be https://doi.org/10.5066/P9DKA9KQ

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

For the updated 2 panel plots, we now stack the plots vertically.

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

For the updated 2 panel plots, we now stack the plots vertically.

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

For the updated 2 panel plots, we now stack the plots vertically.