Author Response to Editor and Reviewer

Editor Comments:
Dear authors,
After subsequent feedback from the reviewers, I'd like to recommend moderate revisions to your manuscript. I strongly encourage you to consider the major recommendation from the reviewer for incorporation into your manuscript.

Overall, the reviewers agree that the manuscript has improved. The reviewers also indicate that there are still changes that can be made to elevate the impact of this study.

I look forward to receiving and reviewing your updated manuscript.

Author Response to Editor:
Dear Editor,
We appreciate the opportunity to revise our manuscript. We have addressed Referee #1's major recommendation, which is to apply the R package from the Clark et al. (2021) paper to compute the KGE uncertainty for each of the gages. To demonstrate the uncertainty in the KGE estimates, we have generated a new figure which we include in the Supplemental Material. Further, the generated KGE uncertainty data for all of the gages was added to each of the respective data releases: for NWMv2.1 they have been added to Towler et al. 2023a and for NHMv1.0 they have been added to Towler et al. 2023b. Please see details in our response to Referee #1, below.

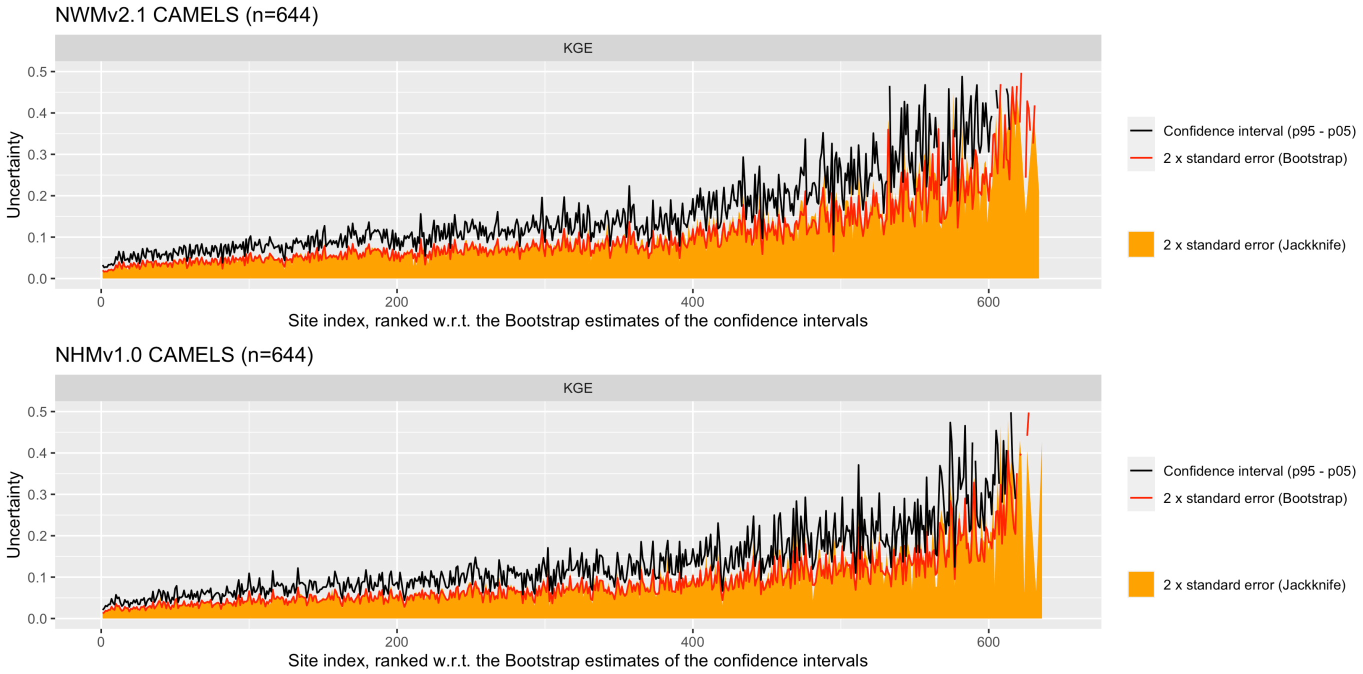
Anonymous Referee #1 Report:
Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)
I would like to thank the authors for their substantial overhaul of this manuscript in response to reviewer suggestions. I think these changes have made the paper more relevant, particularly due to the benchmarking approach the authors applied. I think there is value in publishing this paper as a means to encourage more detailed benchmarking of model performance and to encourage a shift in community thinking away from "look at how high my efficiency scores are" and towards deliberate assessment of model weakness (and hence areas of possible model improvement).

If I have one major comment to make it is that the authors note that "one limitation of this study is that it does not consider the sensitivity of the KGE to sampling uncertainty, which can be large for heavy-tailed streamflow errors (Clark et al., 2021)". This would be a straightforward limitation to address, because this Clark et al. paper also points to an R package that can be used to compute the KGE uncertainty in an easy manner. I believe this would lift the paper to a higher level.

Thank you for the feedback, and we have addressed your major comment. Following from Clark et al. (2021), we have run the gumboot package (Clark and Shook 2021) for all the gages to compute the KGE uncertainty. Using the KGE uncertainty outputs, we have generated a new figure which we include in the Supplemental Material (see later in this response for more details on this figure). Further, the KGE uncertainty data was added to the existing data releases: for NWMv2.1 they have been added to Towler et al. 2023a and for NHMv1.0 they have been added to Towler et al. 2023b. These have been added as individual csv files (with their own metadata); they include the outputs from the gumboot *bootjack()* function for the KGE, including the standard error of jacknife, standard error of bootstrap, the 5th, 50th and 95th percentiles of the estimates, the jackknife score, the bias of jackknife, the bias of bootstap, the standard error of jackknife after bootstrap (Clark and Shook 2021). Each new csv file includes all 5390 gages from Foks et al. (2022), but includes NAs where there is not sufficient data to compute the *bootjack()* function. For the NHMv1, uncertainty estimates could be computed for 5312 out of 5390 gages, and for the NWMv2.1, 5288 of the 5390 gages. Following from Clark et al. (2021), the uncertainty estimates of the KGE estimates were plotted for the CAMELS dataset (Addor et al. 2017) from each model using the gumboot *ggplot_estimate_uncertainties()* function. The figure has been added as Figure 11 of the Supplemental Material. The figure shows that the bootstrap and jackknife yield similar estimates, and that there is uncertainty in KGE for both models, similar to what is found in Clark et al. (2021). We have amended the Discussion as follows:

"Clark et al. (2021) point out that it is important to characterize the sensitivity of KGE to sampling uncertainty, which can be large for heavy-tailed streamflow errors. Using bootstrap methods (Clark et al. 2021), uncertainty in the KGE estimates for this study were computed (Towler et al. 2023a, 2023b) and are illustrated in Supplemental Figure 11."

The below figure and caption were added to the Supplemental Material:


NWMv2.1 CAMELS (n=644)

NHMv1.0 CAMELS (n=644)

Supplemental Figure 11. Estimates of uncertainty in the KGE estimates for the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) basins (Addor et al. 2017) using the gumboot package (Clark and Shook, 2021) in R (R Core Team, 2021) for the National Water Model v2.1 (NWMv2.1; top) and National Hydrologic Model v1.0 (NHMv1.0; bottom). Quantification of the uncertainty is obtained from 2x standard error estimates obtained using jackknife and bootstrap estimates, as well as intervals computed as the difference between the 95th and 5th percentiles of the bootstrap samples (see Clark et al. 2001 for details). The figure shows the uncertainty in the KGE estimates, with the bootstrap and jackknife showing similar estimates for both models. KGE uncertainty estimates for the full set of gages in this study (Foks et al. 2022) are included in Towler et al. (2023a, 2023b).

References:

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.

Clark M. and Shook, K: Package ‘gumboot: Bootstrap Analyses of Sampling Uncertainty in Goodness-of-Fit Statistics’, available online at: https://cran.r-project.org/web/packages/gumboot/index.html (last access March 6, 2023), 2021.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al.: The abuse of popular performance metrics in hydrologic modeling, Water Resour. Res., 57, e2020WR029001. https://doi.org/10.1029/2020WR029001, 2021.

Foks, S.S., Towler, E., Hodson, T.O., Bock, A.R., Dickinson, J.E., Dugger, A.L., Dunne, K.A., Essaid, H.I., Miles, K.A., Over, T.M., Penn, C.A., Russell, A.M., Saxe, S.W., and Simeone, C.E.: Streamflow benchmark locations for conterminous United States, version 1.0 (cobalt gages): U.S. Geological Survey data release, https://doi.org/10.5066/P972P42Z, 2022.

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.r-project.org (last access: 4 May 2022), 2021.

Towler, E., Foks, S.S., Staub, L.E., Dickinson, J.E., Dugger, A.L., Essaid, H.I., Gochis, D., Hodson, T.O., Viger, R.J., and Zhang, Y.: Daily streamflow performance benchmark defined by the standard statistical suite (v1.0) for the National Water Model Retrospective (v2.1) at benchmark streamflow locations for the conterminous United States (ver 3.0, March 2023): U.S. Geological Survey data release, https://doi.org/10.5066/P9QT1KV7, 2023a.

Towler, E., Foks, S.S., Staub, L.E., Dickinson, J.E., Dugger, A.L., Essaid, H.I., Gochis, D., Hodson, T.O., Viger, R.J., and Zhang, Y.: Daily streamflow performance benchmark defined by the standard statistical suite (v1.0) for the National Hydrologic Model application of the Precipitation-Runoff Modeling System (v1 byObs Muskingum) at benchmark streamflow locations for the conterminous United States (ver 3.0, March 2023): U.S. Geological Survey data release, https://doi.org/10.5066/P9DKA9KQ, 2023b.

Beyond that I only have a handful of minor editorial suggestions:

Line 157. "(Falcone 2011)" - There's a comma missing here.
This has been added.
Line 197. "select KGE scores" - Out of curiosity, what prompted the choice of -0.06 (and 0.50 and 0.75) as thresholds? The reasoning for this might be added to the text.
The reason for the -0.06 is already present in the text, but we now specify that one should look at Figure 2 (the KGE CDF plot) to see that -0.06 is where the KGE CDF curves intersect. We have added (change in bold): "**From Figure 2, it can be seen that t**~~T~~he CDFs for the models intersect with the AvgDOY curve at a KGE score of about -0.06; at this value, 19%-20% of the sites perform worse in terms of KGE using the model simulation, whereas above this value the model simulations perform better than AvgDOY." The other values (0.50 and 0.75) were selected to provide select quantitative values to go with the figure.
Line 208. "it" - consider replacing with "a gauge".
This has been added.
Line 227. "Southeast" - It may be good to specify which states area meant here. Looking at Florida, performance seems uniformly quite poor.
We have made the following changes (additions in bold): "Relatively good performance is seen in **most of** the Southeast, **but performance tends to be poor or mixed in Florida**."
Line 247. "you look" - consider replacing with "one looks".
This has been changed.
Line 295-297. "Lane et al. ... in the models." - This sentence might be better moved to Line 289, before "One likely reason ...", with "One likely reason ..." replaced by "This is a likely

explanation in our case as well, because water withdrawal ... ". I think this would improve the flow of this section.

We have incorporated these suggestions.

Table 5. I would suggest to use 1 decimal place for all percentages listed in column "n (%)" so that the sums all cleanly add to 100%.

This has been done.

Figure 3. It's interesting that the simple data-based benchmark model has virtually identical performance in Reference and Non-reference gages, whereas the models clearly struggle a lot more with the Non-reference gages than the Reference gages. This may indicate that, even though many gages are managed, their flows are (generally speaking) about as seasonally stable, and hence about as predictable by this simple model, as the flows at unmanaged gages. This may point to an opportunity to use really simple post-processing tools to simulate water management in the Non-ref basins. No real comment beyond this or any need for the authors to do something with this, I just found it interesting to mention.

Agreed it is an interesting finding.

Figure 5, 6. Is there a particular reason to set the 5 color bins in these plots (and possibly others) at [-Inf, 0.2], [0.2, 0.4], etc.? Having a lower bin of [-Inf, -0.41] might make more sense, as this gives some reason to lump everything below those numbers together. The remaining 4 bins might be evenly divided as [-0.41, -0.06], [-0.05, 0.30], etc.

We have changed the color bins for Figure 5 and 6 as suggested; see below:

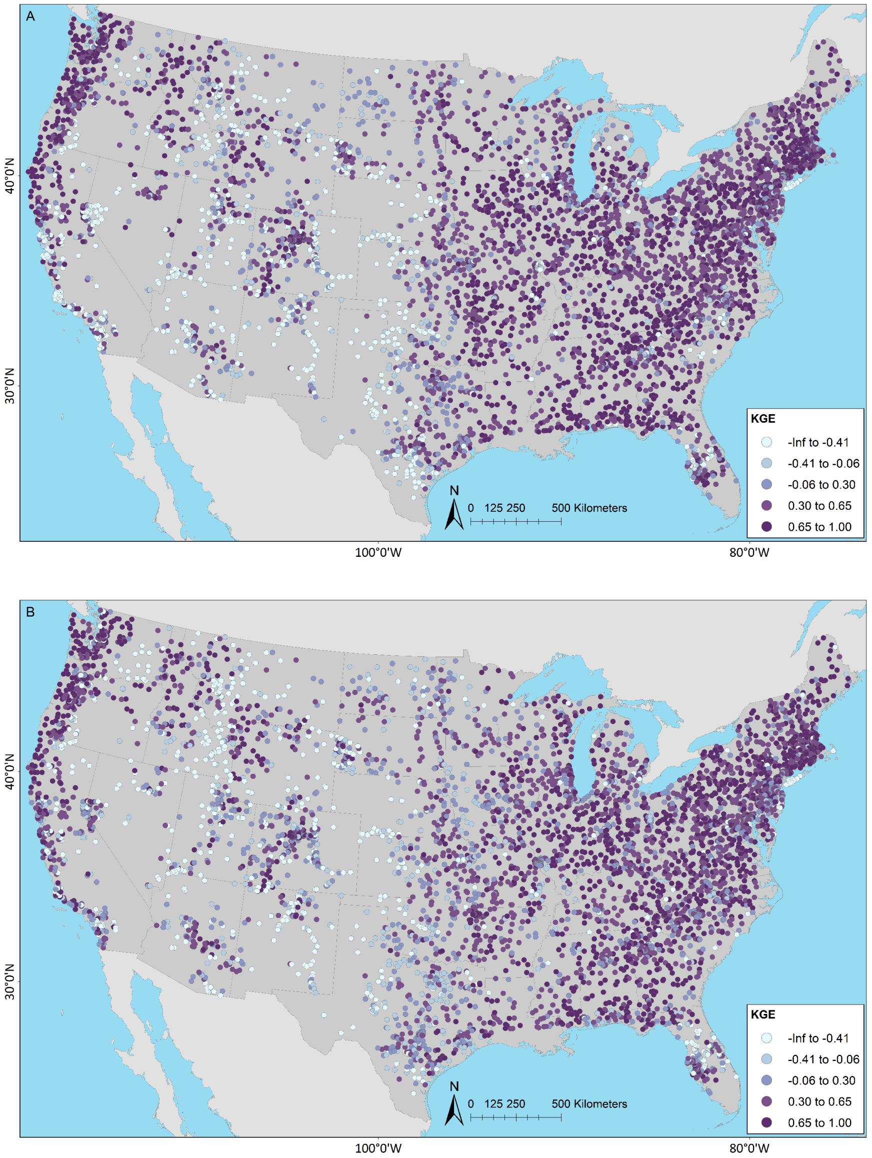


Figure 5: Kling–Gupta efficiency (KGE) based on daily streamflow at U.S. Geological Survey (USGS) gages for (A) National Water Model v2.1 (NWMv2.1) and (B) National Hydrologic Model v1.0 (NHMv1.0). The Map Source: (Grannemann, 2010; Natural Earth Data, 2009; ESRI, 2022a; ESRI, 2022b).

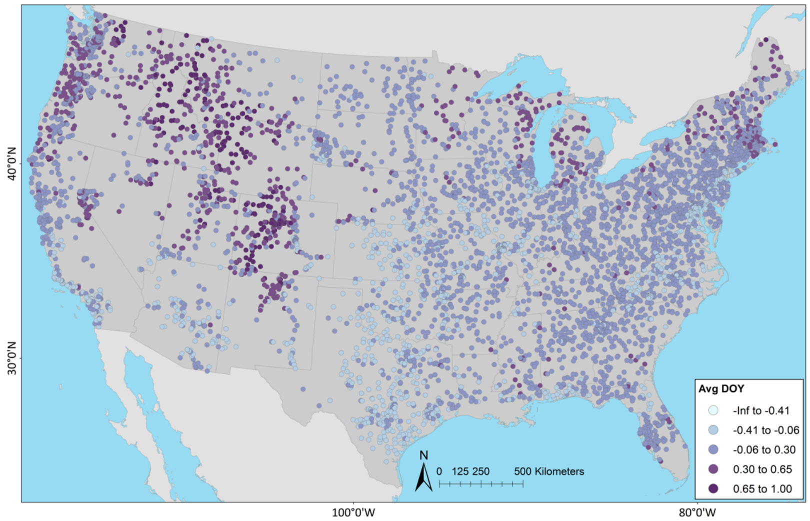


Figure 6: Kling–Gupta efficiency (KGE) based on daily streamflow at U.S. Geological Survey (USGS) gages using seasonal benchmark from average day-of-year flows (AvgDOY). Map Source: (Grannemann, 2010; Natural Earth Data, 2009; ESRI, 2022a; ESRI, 2022b).