

Author response

Dear editor, dear reviewers,

Thank you for your assessment of our work. Your comments have been very helpful in clarifying the main message we hope to convey. Please find our responses to your individual comments below. These are either the same or expanded versions of our initial response to your comments.

In addition to the changes described below, we've made a number of minor clarifications and corrections to the text. These do not affect the main conclusions but can be viewed in the track-changes document as well. Please note that the new Table 1 runs over the edge of the page in the track-changes version of the document but renders properly in the revised manuscript.

On behalf of all co-authors,

Kind regards,

Wouter Knoben

Response to reviewer 1

The authors provide a nice and easy-to-read study on large-domain hydrological model evaluation. They introduced benchmarks as a valuable tool to investigate where a large domain hydrologic model might still lack in performance. This makes the manuscript, in my opinion, very interesting for a wider hydrological audience. Before publication, however, I think the manuscript would benefit from a bit more analysis on why the model is failing in certain areas. Below, I try to give some constructive feedback for the authors to consider:

Thank you for your comments. It has been very helpful to understand which parts of our reasoning you think should be clarified. Please see our responses to your individual comments below.

General remarks:

- I have the feeling that a simple explanation of why a model fails if it is worse than a certain benchmark would help readers to understand the point of this technical note.

We have restructured the introduction to more sharply focus on the benchmarking problem our technical note addresses and include an example using the Nash-Sutcliffe Efficiency. Most readers will be familiar with the NSE, and it can therefore show how and why comparing a model to a benchmark is useful (because that is one possible way of interpreting the NSE), as well as highlight the need to improve our benchmarking (because the benchmark included in NSE does not impose strong constraints on model performance):

“The deliberate use of benchmarks can provide a helpful frame of reference for interpreting efficiency scores such as NSE and KGE, by setting realistic expectations of the possible performance in each basin (Seibert, 2001; Schaefli and Gupta, 2007; Legates and McCabe, 2013; Seibert et al., 2018; Beven, 2023; Knoben, 2024). A well-known example follows from a specific interpretation of the Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970):

$$NSE = 1 - \frac{\sum_{t=1}^N (q_{obs}(t) - q_{sim}(t))^2}{\sum_{t=1}^N (q_{obs}(t) - \bar{q}_{obs})^2},$$

where q_{obs} and q_{sim} are observed and simulated streamflow respectively. This equation can be interpreted as a skill score that quantifies how much of the variance in q_{obs} the model (through q_{sim}) explains compared to the simple benchmark that is the long-term mean flow \bar{q}_{obs} . Although this specific benchmark is often criticized for the limited constraints it imposes on model performance (e.g. Schaefli and Gupta, 2007), it provides a useful example of a simple benchmark. By comparing the performance of a model against a (much) simpler alternative way of predicting the variable of interest, it becomes easier to evaluate if and how much better the hydrologic model is.”

We’ve also considerably expanded Section 2.4 (where the benchmarks are introduced) and made a number of targeted changes to the Results section, to ensure that the main point we wish to make (benchmarking reveals areas for model improvement that might be missed with traditional metrics) comes through clearly.

- I have the feeling the Introduction is not well linked to the rest of the manuscript. I did not get the impression that the questions raised were answered. The manuscript does not give any guidance if a score is indicative or useful for a model, nor does it go into quantifying their uncertainty. Isn’t the point of the manuscript more to find regions and reasons where the model is failing against the suggested benchmarks? I recommend restructuring the introduction accordingly.

Thank you for bringing this up. We’ve rewritten the introduction with a sharper focus on our main question. Key changes:

- Replace the current

“To effectively use these model performance statistics, two questions need to be answered. First, is this score indicative of a useful model for the purpose at hand? Second, how uncertain is this score?”

text at the end of paragraph 2 (line 26-28) with:

“These scores are useful because the community has relied on them for a long time and they now function as an informal shared test environment (Clark et al., 2026). However, a key challenge remains that the scores calculated by these metrics are difficult to interpret in isolation (e.g., Seibert, 2001; Schaefli and Gupta, 2007; Knoben, 2024), partly

because they tend to conflate model performance and flow variability (Schaefli and Gupta, 2007; Williams, 2025; Clark et al., 2026).”

This text is then immediately followed by the changes proposed in our previous response.

- Immediately after the NSE example, we’ve added a new paragraph that condenses the current text to increase focus and provide extra support for the use of simple benchmarks:

“Benchmarks can take various forms, such as regression equations (as used in certain land modeling experiments; e.g., Best et al., 2015), statistics such as persistence or climatology (as common in the streamflow forecasting community; e.g., Harrigan et al., 2023), or different versions of the same model (to see if model changes have the desired effect; e.g., Cosgrove et al., 2024). Benchmarking is also commonly seen when models of varying levels of complexity are compared, particularly in current large-domain modeling exercises that contrast the performance of machine learning methods to more traditional hydrologic models (e.g., Kratzert et al., 2019; Song et al., 2025). The main trade-off between different types of benchmarks is the cost of employing the benchmarks compared to what can be learned from them. For example, the cost of comparing an existing hydrologic model against a second one is often prohibitive because configuration is too cumbersome, or run times too long, but comparing the performance of any model against a simple baseline has been common practice as long as the Nash Sutcliffe Efficiency has been in use. Using simple benchmark models, such as the long-term mean, gives some idea of the predictability of the streamflow observations in each basin at negligible computational cost. Our hypothesis is that comparing the performance of a model against the performance of an ensemble of simple benchmarks can be an effective way to identify cases where the performance of a large-domain model is not as high as it could be, irrespective of the absolute values of the scores, and thus where opportunities for model improvement may exist.”

This is then followed by a paragraph on sampling uncertainty and the ending paragraph of the introduction.

We do note that quantifying the (sampling) uncertainty of the scores is plays a main role in the manuscript – it just matters less in this case than in some other applications. We have added histograms to Figure 3 to show how often $J == 0$ (i.e., sampling uncertainty plays no role). In our main analysis, this is the case in roughly three quarter of the basins.

We have made a number of text changes to Section 3.3 (lines 346-353 in the track-changes document) and added a specific sentence on lines 359-360 to report precise numbers.

- Discussion: I would like to see a more in-depth discussion on what it actually means if the Benchmark is better than the model. After that, you can go into the analysis, where and why the model might have failed. For this, however, I would recommend putting more emphasis on why the model has failed. Maybe look at it from a model development perspective, what would you need to do to improve the model? Try to give some guidance. E.g., by correlating your J index against catchment attributes (soil, landuse,

climate, etc.), and also against the KGE, might give more insights. I acknowledge that the authors already provide some discussion on why the model might fail under certain circumstances, but very few of them are really based on the results of this manuscript and are rather based on the authors' knowledge and other literature.

Hopefully the changes to the earlier parts of the text more clearly identify the main point of our paper: that benchmarking reveals areas for model improvement that might be missed with traditional metrics. The NWM retrospective simulations provide a readily available large-domain data set originating from a model that is used operationally, and are therefore a useful test case to show the merits of using benchmarks. Identifying the specific reasons for why the NWM fails is beyond the scope of this work: it would require much more in-depth technical evaluation than is feasible to add to this paper, it should be done with the operational version of the model (not the daily-aggregated retrospective simulations) to be practically useful, and – most importantly – detract from the main point that benchmarking can be a valuable tool in any modeling work, not just this specific one involving the NWM.

Minor comments:

- Title: Is it really fitting what the manuscript is about? I would suggest something like “Technical note: Benchmarking large-domain hydrological model performance”. If the authors want to state uncertainty in their title, they should be specific what kind of uncertainty they are referring to.

We have changed the title to “Technical note: Benchmarking large-domain model performance under sampling uncertainty”.

(Was: Technical note: Separating signal from noise in large-domain hydrologic model evaluation - Benchmarking model performance under sampling uncertainty

We note that “sampling uncertainty” is the commonly used phrase for the uncertainty stemming from the choice of data used to calculate any model performance metric on. We acknowledge the reviewer’s concern that readers may not be familiar with the term however (it is a relatively recent development after all) and have updated the abstract to clarify this for the reader (addition in bold):

“Abstract. Large-domain hydrologic modeling studies are becoming increasingly common. The evaluation of the resulting models is however often limited to the use of aggregated performance scores that show where model accuracy is higher and lower. Moreover, the inherent uncertainty in such scores (**i.e., the sampling uncertainty**), stemming from the choice of time periods used for their calculation, often remains unaccounted for.”

- Introduction: What are simple benchmarks exactly, where have they been used, and what's their benefit, how do they relate to hydrological signatures?

The changes to the introduction should address part of this comment already (what are simple benchmarks and what is the benefit of using them?). We have extended and rewritten Section 2.4

where the specific benchmarks are introduced to briefly address where these benchmarks have been used and to note why we use a benchmark ensemble:

“Hydrologic models are increasingly compared to more taxing benchmarks than the long-term mean flow (e.g. Knoben et al., 2020; Towler et al., 2023), but outside the forecasting community (see e.g. Pappenberger et al., 2015) such work is still somewhat limited. Benchmarks also vary in their strengths and weaknesses, and what constitutes a strong benchmark can change regionally (Pappenberger et al., 2015; Knoben, 2024). We therefore compare the performance of the NWM to the performance of an ensemble of simple benchmark models that cover various levels of complexity. A full list of the 17 different benchmark models used in this work can be found in Table 1. These benchmarks are effectively an “ensemble of opportunity”: they are conveniently available in the HydroBM package (Knoben, 2024) and serve to illustrate the point made in the remainder of this paper. We note that this benchmark ensemble is neither exhaustive, nor is it meant to be. However, as long as more theory-driven benchmark selection methods are lacking (i.e., selecting a specific benchmark for a specific basin, based on the benchmark’s suitability for representing the basin’s specific flow regime), ensemble benchmarking methods provide an acceptable alternative.”

There is no immediate relation between the benchmarks and streamflow signatures, and diving into this is a bit beyond the scope of this manuscript.

- Section 2.4 Benchmarks: It should be better explained which benchmarks are actually used and why.

The benchmarks were originally explained in more detail in the Supporting Information, but in hindsight perhaps this information is better placed in the main body of the paper. Reviewer 2 specifically requests this as well. We’ve moved the information into the main paper and clarified our reasoning for choosing the ones we used. The following was added to Section 2.4:

“A full list of the 17 different benchmark models used in this work can be found in Table 1. These benchmarks are effectively an “ensemble of opportunity”: they are conveniently available in the HydroBM package and serve to illustrate the point made in the remainder of this paper. We note that this benchmark ensemble is neither exhaustive, nor is it meant to be. However, as long as more theory-driven benchmark selection methods are lacking (i.e., selecting a specific benchmark for a specific basin, based on the benchmark’s suitability for representing the basin’s specific flow regime), ensemble benchmarking methods provide an acceptable alternative (Knoben, 2024).”

We’ve also added a new subsection to Section 2.4 where we describe the benchmarks in more detail than we did previously:

“Each benchmark represents a simple way of predicting the variable of interest (here: streamflow), and thus sets a certain minimum expectation of how well a specific aspect of catchment behavior can be predicted. This in turn can be seen as a test for the model of interest: if the model underperforms compared to the simple alternative, improvements to

the modeling chain may be possible. For example, if a model shows consistent bias during low flows but a simple seasonal cycle benchmark does not, this suggests that the flows themselves are relatively stable between years but that the model is somehow unable to replicate this pattern. The benchmark does not immediately point out the underlying causes of the model's bias, but it does show that model performance is not as high as it can be. As shown in Table 1, the benchmarks cover three different categories.

The first category covers simple statistics calculated from the streamflow observations, which are then used as a predictor of streamflow on all time steps. These benchmarks thus quantify the stability of the flow regime in time by using past observations to provide an estimate of how flows at any given point in the future might look, and thus challenge the model to predict deviations from the catchment's typical streamflow behaviour. One example is the long-term mean flow which, if used as a predictor of flow, returns a time series of constant values. This benchmark appears as part of the denominator in the Nash-Sutcliffe efficiency and has been commonly used in hydrology, although often criticized for the limited constraints it imposes on model performance (e.g., Schaefli & Gupta, 2007) outperforming this benchmark is, in most cases, not very difficult. A second example is the daily mean flow which characterizes the typical seasonal cycle of the flow regime. If the flow in any given year is different from the typical seasonal regime, the model should be able to predict these deviations. If it does, its performance will be higher than the benchmark's.

The second category covers benchmarks that attempt to account for the influence of precipitation on streamflow. These benchmarks first calculate the average rainfall-runoff ratio (or ratios, in the case of the monthly benchmarks), and then use this ratio to scale incoming precipitation. This approach assumes that the amount of precipitation influences a catchment's streamflow response, but that the ratio of precipitation-to-streamflow conversion does not change markedly throughout time. These benchmarks thus challenge the model to predict deviations from typical rainfall-runoff ratios, which may be the case under prolonged drying or anomalous wet conditions. An example is the benchmark that applies average monthly rainfall runoff ratios to monthly precipitation totals. Despite its coarse temporal resolution (flows within a month are constant), this benchmark has shown considerable performance in a previous large-domain application (Knoben, 2024).

The benchmarks in the third and most complex category are still rather simple one- and two-parameter models whose parameters are optimized using a brute-force approach. These benchmarks attempt to capture the main components of catchment behavior (i.e., partitioning, delayed response, attenuation of precipitation inputs) in parsimonious and aggregated ways. This approach challenges the model to see if the addition of further degrees of freedom (i.e., having more parameters) leads to an appreciable increase in predictive performance. The most complex benchmark in this category is the two-parameter Adjusted Smoothed Precipitation Benchmark (ASPB) proposed by (Schaefli & Gupta, 2007). This benchmark scales incoming precipitation by the long-term rainfall-runoff ratio to simulate precipitation partitioning, smooths the resulting scaled precipitation with a moving window approach of calibrated length, and then shifts this smoothed

response by a calibrated lag value. This provides a two-parameter approximation of the main components of catchment behaviour.”

- Figure 2: Might it be easier to focus on the evaluation period only? And maybe I missed it in the Data and Methods section, but it should be clearly defined what the evaluation period is. Section 2.4 is speaking of a validation period; is this used as a synonym here? If so, it would be better to use only one of the two words throughout the manuscript.

There are three points here:

1. Should Figure 2 focus on the evaluation period only?

There is value in showing both calibration and evaluation performance of the model and the benchmarks. Calibration performance shows data fitting potential (i.e., how well can a given method – model or benchmark – capture the data at all in a given basin?), whereas evaluation performance shows what sort of predictive power that data fit actually has (i.e., how well can does a given method – model or benchmark – capture underlying processes in a way that’s transferable in time?). We’ve added the following to the start of Section 3.1:

“Calibration performance shows data fitting potential (i.e., how well can a given method – model or benchmark – capture the patterns in the data at all in a given basin?). Evaluation performance shows what sort of predictive power that data fit actually has (i.e., how well can a given method capture the underlying processes in a way that leads to accurate predictions for unseen data?)”

and the following to the next paragraph where these results are discussed:

“For calibration, this suggests that the NWM (14 calibrated parameters in NWMv2.1, (Cosgrove et al., 2024), assumed to be a similar number for the NWMv3.0 calibrations shown here), as may be expected, has greater flexibility than the benchmarks (0 to 2 parameters) to fit to the specific characteristics of the calibration data. For evaluation, the CDFs of both model and benchmark performances show a tendency towards lower scores. This is commonly seen in any modeling study and typically attributed to some degree of overfitting of the model to specifics of the calibration data, or to a change in conditions between calibration and evaluation periods that the model cannot effectively account for. Some benchmarks (e.g., BM11, Fig. 2k) show very limited performance change, suggesting that they capture the aggregated catchment response equally well (or poorly) during both data periods. Other benchmarks (e.g., BM07, Fig. 2g) show very large performance changes, suggesting that calibration conditions were not sufficient to let the benchmark accurately capture underlying catchment behavior. Compared to the benchmark ensembles, the NWM does not stand out as having particularly large or small performance changes.”

2. What is the evaluation period?

This is described in Section 2.4 Benchmarks (underlined for emphasis here):

“We configure the benchmark models in the same way as a regular model application would be structured: the benchmarks are defined using data from a dedicated calibration period (though “calculation period” is a more accurate description for most benchmarks, because only BM 16 and BM17 require parameter calibration) and then used to predict the streamflow in an independent validation evaluation period. We used the same 5-year time period to calibrate the benchmarks as was used to calibrate the NWMv3.0: from 2016-10-01 to 2021-09-30. In case the observation data were incomplete, we used either 4 or 3 water years within that same 5-year window instead. The validation period is all the data from 1980-01-01 to 2022-12-31 that is not used for calibration.”

To aid the reader, we have added the following text to the section where the results are first introduced (additions in bold):

“Figure 1 shows the KGE scores obtained by the NWM as well as the 17 benchmark models. Performance is shown as Cumulative Distribution Functions (CDFs) for straightforward comparison of performance aggregated across all locations. **Results are shown for both the calibration period (up to water 5 years of data used, depending on data availability at each gauge) and the evaluation period (up to 37 water years).**”

3. Is there a difference between validation and evaluation periods?

“Validation” and “evaluation” are indeed used interchangeably throughout the manuscript. We have replaced all occurrences of “validation” with “evaluation” for consistency.

- Figure 3: Show what BMs are actually standing for; that’s not too much text for the figure.

We agree that this is more mysterious than it needs to be. We’ve added more descriptive titles, as also requested by reviewer 2. The NSE figure in the Supporting Information was updated as well.

Response to reviewer 2

Review of "Technical note: Separating signal from noise in large-domain hydrologic model evaluation - Benchmarking model performance" by Gründemann et al.

The technical note promotes the use of various benchmarks for model performance evaluation, particularly in a large-domain setting (or for large-sample studies) and includes a quantification of sampling uncertainty from different periods through bootstrapping of different hydrological years.

The manuscript is clearly, concisely written and well structured.

Thank you for your comments. They have been very helpful in outlining how the manuscript might be refined. Please see our responses to your individual comments below.

Before I can recommend publication, however, I would like to raise the following comments:

major comments:

- Since this note is all about the benchmarks, I think two ingredients are missing:

1) Please add the benchmarks and their description to the main text and not just to the supplementary material and ensure that the abbreviations match those in the figures (or vice versa)

Reviewer 1 similarly requests more explanation of the benchmarks in the main body of the text. We've moved the information as suggested and corrected any occasions we could find where incorrect benchmark abbreviations were used.

2) Each of the benchmarks is essentially a test of how well a model should minimally perform regarding a specific aspect. This is not discussed in detail in the manuscript, but I think providing some examples would really help promoting the use of various benchmarks from very simple ones targeting maybe the water balance to more complex ones. I would suggest extending the discussion and conclusions accordingly and as well as adding this explanation regarding which aspect they are benchmarking in the table describing them.

We have made the following changes in response to this comment:

- As mentioned above, we've moved the benchmark table from the appendix to the main manuscript
- We've extended the benchmark descriptions in the table, and added broad categorical headers that describe what each category of benchmarks is supposed to test. Combined with the other changes to Section 2.4 described below, we think this is a more space-efficient way of providing the reader with the same information (compared to adding a column to the table). *Please note that in the track-changes document the table runs over the edge of the page – it should be viewable in its entirety in the updated doc that does not highlight the changes made.*
- We've added a dedicated sub-section to Section 2.4 Benchmarks, where we discuss the different benchmark categories, as follows:

“Each benchmark represents a simple way of predicting the variable of interest (here: streamflow), and thus sets a certain minimum expectation of how well a specific aspect of catchment behavior can be predicted. This in turn can be seen as a test for the model of interest: if the model underperforms compared to the simple alternative, improvements to the modeling chain may be possible. For example, if a model shows consistent bias during low flows but a simple seasonal cycle benchmark does not, this suggests that the flows themselves are

relatively stable between years but that the model is somehow unable to replicate this pattern. The benchmark does not immediately point out the underlying causes of the model's bias, but it does show that model performance is not as high as it can be. As shown in Table 1, the benchmarks cover three different categories.

The first category covers simple statistics calculated from the streamflow observations, which are then used as a predictor of streamflow on all time steps. These benchmarks thus quantify the stability of the flow regime in time by using past observations to provide an estimate of how flows at any given point in the future might look, and thus challenge the model to predict deviations from the catchment's typical streamflow behaviour. One example is the long-term mean flow which, if used as a predictor of flow, returns a time series of constant values. This benchmark appears as part of the denominator in the Nash-Sutcliffe efficiency and has been commonly used in hydrology, although often criticized for the limited constraints it imposes on model performance (e.g., Schaepli & Gupta, 2007) outperforming this benchmark is, in most cases, not very difficult. A second example is the daily mean flow which characterizes the typical seasonal cycle of the flow regime. If the flow in any given year is different from the typical seasonal regime, the model should be able to predict these deviations. If it does, its performance will be higher than the benchmark's.

The second category covers benchmarks that attempt to account for the influence of precipitation on streamflow. These benchmarks first calculate the average rainfall-runoff ratio (or ratios, in the case of the monthly benchmarks), and then use this ratio to scale incoming precipitation. This approach assumes that the amount of precipitation influences a catchment's streamflow response, but that the ratio of precipitation-to-streamflow conversion does not change markedly throughout time. These benchmarks thus challenge the model to predict deviations from typical rainfall-runoff ratios, which may be the case under prolonged drying or anomalous wet conditions. An example is the benchmark that applies average monthly rainfall runoff ratios to monthly precipitation totals. Despite its coarse temporal resolution (flows within a month are constant), this benchmark has shown considerable performance in a previous large-domain application (Knoben, 2024).

The benchmarks in the third and most complex category are still rather simple one- and two-parameter models whose parameters are optimized using a brute-force approach. These benchmarks attempt to capture the main components of catchment behavior (i.e., partitioning, delayed response, attenuation of precipitation inputs) in parsimonious and aggregated ways. This approach challenges the model to see if the addition of further degrees of freedom (i.e., having more parameters) leads to an appreciable increase in predictive performance. The most complex benchmark in this category is the two-parameter Adjusted Smoothed Precipitation Benchmark (ASPB) proposed by (Schaepli & Gupta, 2007). This benchmark scales incoming precipitation by the long-term rainfall-runoff ratio to simulate precipitation partitioning, smooths the resulting scaled precipitation with a moving window approach of calibrated length, and then shifts this smoothed response by a calibrated lag value. This provides a two-parameter approximation of the main components of catchment behaviour.”

- there is the sampling uncertainty, there is the model uncertainty, but what makes up these metrics are also affected by the uncertainty inherent in the observations. It would be worth

reminding the reader that these can be considerably large and influential on the performance metric. For instance, for discharge, there is the rating curve uncertainty that is not constant but varies with the flows (see for instance Westerberg et al., 2011)

We've added this reminder to the introduction, where sampling uncertainty is first introduced (addition in bold)

“However, assessing if a model outperforms a benchmark is not always straightforward. **Even if ignoring the fact that observational uncertainty may mean that model simulations are being compared to incorrect data (e.g. Westerberg et al., 2011; Gharari et al., 2024)**, a confounding issue is that performance scores such as NSE and KGE are inherently conditional on the time period for which they are calculated (McCuen et al., 2006; Ritter and Muñoz-Carpena, 2013; Lamontagne et al., 2020; Clark et al., 2021; Klotz et al., 2024).”

Line by line comments:

Abstract

L4 name at least some examples of what is meant by a simple benchmark, i.e. make it more specific

We've added a sentence to the abstract containing an example of each of the three main benchmark categories that we recognize:

“These benchmarks are simple ways of predicting the variable of interest (here, streamflow) and include, for example, the long-term daily mean flow, daily precipitation scaled by the average rainfall-runoff ratio, and a basic 2-parameter model that represents a catchment's diffusive response to precipitation inputs.”

L5-7 these results are valid for the study region and basins and but not for other regions, please add that the data set is from the United States and maybe add even NWM

We've added another sentence to the abstract to make this explicit:

“Our test case consists of simulations from the National Water Model v3.0 for approximately 4900 basins across the United States.”

We also made a minor change to the text in Section 2.1 that (hopefully) prevents any potential confusion w.r.t. the basin number mentioned there. “3800 locations” simply refers to gauges where the River Forecast Centers provide forecasts, which is something separate from the National Water Model simulations at the 4,900 gauges we use in this work.

L9 ", though accounting..." this part of the sentence is not clear. Please rephrase.

We'll rephrase as follows (changes in bold):

“Sampling uncertainty has limited impact: in most basins the model is either clearly better or worse than the benchmarks, **though numerous cases remain where sampling uncertainty makes it difficult to clearly distinguish between model and benchmark performance.**”

(was: “Sampling uncertainty has limited impact: in most basins the model is either clearly better or worse than the benchmarks, though accounting for sampling uncertainty remains important when the performance of different models is more similar.”)

Main text

L21-25 the words "score", "statistics", "efficiency", "metrics" are used and they are used interchangeably. I would suggest using only one, where this is applicable and using it consistently throughout the manuscript

Thanks for pointing this out. This is particularly important because we use “statistics” as “statistical moments calculated from the streamflow time series” (e.g., the mean) and this needs to be kept separate from “efficiency scores” and “metrics”. We’ve cleaned up the use of terms throughout the text and add a few sentences on definitions at the end of the introduction:

“In the remainder of the text, we use the following definitions:

- *Statistics*: summary statistics derived from a time series (e.g., the long-term mean of flow observations, the daily median flow).
- *Metrics*: specific equations used to summarize model performance into a single number (e.g., the Root Mean Squared Error, the Nash-Sutcliffe efficiency).
- *Performance scores*: values found for a given metric (e.g., the distribution of KGE values obtained when calibrating a given model for a set of basins).”

L22 "and more" remove (there is already "for example" in the same sentence) - changed, and an extra “and” added.

L34 ... or further checks are required

Valid point. However, we’ve rewritten the introduction and this sentence is no longer present.

L40 "can be " -> "is" – changed.

L120 since the benchmarks are the core of this note, Table S1 should be moved to the main text and the abbreviations adjusted accordingly. – please see our earlier answer.

L126 "as" -> "that" – we believe “as” being grammatically correct here, as short for “We used the same 5-year time period to calibrate the benchmarks as [the period that] was used to calibrate the NWM.”

L239 Supporting – changed, thanks

L239 abbreviation was already introduced in L25 - [changed](#)

L257 "perform" missing? – perhaps not, but “performs” sounds better than “does”. [Changed](#).

L262 which benchmark? please add – [this uses a daily mean flow benchmark \(our BM7\). We’ve added this.](#)

L284 remove "and" before "snow" – [changed](#).

Figure 2 in the upper panel the lines are not distinguishable in b&w print

[We’ve replaced this figure with a 6x3 panel that compares the NWM to each benchmark individually. Lines should be clear in black and white.](#)

Figure 3 Please add the written-out benchmarks in the caption so that the figure can stand-alone.

[We’ve added more descriptive titles, as also requested by reviewer 1. The NSE figure in the Supporting Information was updated as well.](#)

References

Westerberg, I., Guerrero, J. L., Seibert, J., Beven, K. J., & Halldin, S. (2011). Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25(4), 603-613.