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'll move the information as suggested, and ensure the abbreviations match the figures.

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 propose the following changes in response to this comment:

- As mentioned above, we'll move the benchmark table from the appendix to the main manuscript
- We'll extend the benchmark descriptions in the table, and add broad categorical headers that describe what each category of benchmarks is supposed to test. Combined with the other changes to Section 2.4 proposed 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).

- We'll 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., 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."

- 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'll add 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 Stated and maybe add even NWM

We'll add another sentence to the abstract to make this explicit:

> "Our test case consists of simulations from the National Water Model v3.0 for approximately 4800 basins across the United States."

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'll clean 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, in the proposed rewrite of the introduction 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 (our BM7). We’ll add this.

L284 remove "and" before "snow" – changed.

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

We’ll trial a 6x3 orientation of subplots with each subplot containing a comparison of just a single benchmark and the NWM, and decide if the increased ability to distinguish individual benchmarks compensates for the lack of ability to compare the performance of the benchmark ensemble (i.e., all of them together in a single plot) against that of the NWM.

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

We’ll add more descriptive titles, as also requested by reviewer 1. The NSE figure in the Supporting Information will be 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.