https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

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

Gaby J. Gründemann[1], Wouter J. M. Knoben[1], Yalan Song[2], Katie van Werkhoven[3], and Martyn P. Clark[1]

[1]Schulich School of Engineering, University of Calgary, Alberta, Canada
[2]Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, United States of America
[3]Research Triangle Institute, Research Triangle Park, North Carolina, United States of America

Correspondence: Wouter J. M. Knoben (wouter.knoben@ucalgary.ca)

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, stemming from the choice of time periods used for their calculation, often remains unaccounted for. Here we use a collection of simple benchmarks whilst accounting for this sampling uncertainty to provide context for the performance scores of a large-domain hydrologic model. The benchmarks suggest that there are considerable constraints on the model's performance in approximately one-third of the basins used for model calibration and in approximately half of the basins where model parameters are regionalized. 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. The areas where the benchmarks outperform the model only partially overlap with areas where the model achieves lower performance scores, and this suggests that improvements may be possible in more regions than a first glance at model performance values may indicate. A key advantage of using these benchmarks is that they are easy and fast to compute, particularly compared to the cost of configuring and running the model. This makes benchmarking a valuable tool that can complement more detailed model evaluation techniques by quickly identifying areas that should be investigated more thoroughly.

1 Introduction

There is a pressing societal need for predictions of water-related risks across large geographical domains. Consequently, water resources modeling at national, continental and global scales is becoming increasingly common (e.g., Arheimer et al., 2020; Cosgrove et al., 2024; Nearing et al., 2024; Song et al., 2025; Van Jaarsveld et al., 2025). Thorough evaluation of such large-domain models is a necessity to improve our understanding of the water cycle, our ability to model it accurately, and to ensure the usability and reliability of model simulations for decision making.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

There is considerable guidance on model evaluation, focusing for example on diagnostics (e.g., Gupta et al., 2008, 2012), multi-variate evaluation (Rakovec et al., 2016; Döll et al., 2024, e.g.,), multi-objective evaluation (e.g., Efstratiadis and Koutsoyiannis, 2010; Kollat et al., 2012), and more. A common theme between these different approaches to model evaluation is that model performance tends to be quantified through performance metrics such as the Root Mean Squared Error (RMSE), the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) and the Kling-Gupta efficiency (KGE; Gupta et al., 2009). Such metrics summarize the (mis)match between observations and a model's simulations as a single performance score. 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?

The first question can be addressed by benchmarking the performance of the model against the performance of other methods that can generate the variable of interest. The goal of the benchmarks is to set realistic expectations about the possible performance in each basin (Seibert, 2001; Schaefli and Gupta, 2007; Legates and Mccabe, 2013; Seibert et al., 2018; Beven, 2023; Knoben, 2024). In a practical context, benchmarks are also helpful to judge the value of investing in a new approach by answering the question - "can our model beat a cheap (i.e., quick and easy to produce) or existing option?". If the answer is "no", then it may not be worth the investment to operationalize the new approach.

Benchmarks typically take the form of other models of varying complexity, and in the case of large-domain hydrologic models could include different versions of the same model, a different hydrologic model, regression equations, and more. The main trade-off between different types of benchmarks is the cost of employing the benchmarks compared to what can be learned from them. Using simple benchmark models for this purpose gives some idea of the predictability of the streamflow observations in each basin at negligible computational cost.

Benchmarking also accounts for the fact that what constitutes a "good value" for scores such as NSE and KGE can be highly variable between basins (Schaefli and Gupta, 2007). For example, Knoben (2024) shows that for various locations across the globe even very simple models might obtain KGE scores as high as $\approx$0.8 when being used to predict unseen data, while for other locations the same models struggle to achieve scores much above the $1-\sqrt{2}$ score that the long-term mean would achieve.

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.

The second question (how uncertain is this score?) addresses the fact 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). Both Clark et al. (2008) and Newman et al. (2015) show that, depending on the nature of the streamflow observations, a large fraction of the total model error may be concentrated in a disproportionally small number of time steps. In such cases, choosing a different period to calculate the scores on might give a very different assessment of the performance of the model. This is commonly referred to as sampling uncertainty. Sampling uncertainty can be considerable (Lamontagne et al., 2020; Clark et al., 2021), and in many cases the scores obtained by different models have uncertainties greater than the differences between them (Clark et al., 2021; Knoben et al., 2025). This complicates

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

the assessment of differences between models, because the models might be statistically indistinguishable. However, the extent to which sampling uncertainty plays a role in large-domain model benchmarking is currently unknown.

In summary, benchmarking is common in the land modeling (e.g., Best et al., 2015) and streamflow forecasting communities (e.g., Harrigan et al., 2023), and is standard practice when multiple versions of the same model are compared (e.g., Cosgrove et al., 2024). It 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). There is, however, limited work on using benchmarks to provide assessments of large-domain predictability of hydrologic response (Seibert et al., 2018; Knoben, 2024), particularly while also considering the effect of sampling uncertainty.

This paper combines the use of simple benchmarks with quantification of sampling uncertainty to evaluate the performance of a large-domain water model, and so demonstrate the value of simple benchmarks in quickly identifying regions where the modeling chain may be improved. In Section 2 we introduce the model (Section 2.1), data (Section 2.2) and performance metric (Section 2.3) used in the analysis, and provide a more in-depth discussion of benchmarks (Section 2.4) and sampling uncertainty (Section 2.5). Results are presented in Section 3, separated into an aggregated assessment of model and benchmark performance (Section 3.1), and a spatial analysis of the results (Section 3.2). We briefly discuss our findings in Section 4 and present our conclusions in Section 5.

2 Data and Methods

2.1 National Water Model v3.0 retrospective simulations

We selected simulations from the National Water Model v3.0 (NWMv3.0) as a practical test case for our work, to investigate our hypothesis that deliberate use of benchmarks can help identify areas for model improvement. The National Water Model is used to generate operational forecasts across the United States, and is primarily designed to produce short-range and medium-range (18 hours to 10 days) sub-daily streamflow forecasts. The NWM forecasts complement those made by the various River Forecast Centres at approximately 3800 locations across the United States by providing forecasts for approximately 3.4 million river reaches. The structure and setup of the NWMv3.0 are similar to those of NWMv2.1 (NOAA, 2025) and are described in more detail in Cosgrove et al. (2024).

We use the NWMv3.0 simulations from the NOAA National Water Model CONUS Retrospective Dataset for the period 1980-01-01 to 2022-12-31. Note that not all gauges have records for the entire period, and in some cases the period of analysis was thus shorter than the full length for which simulations are available. In the retrospective simulations, parameters for the NWM are obtained through a combination of calibration (i.e., parameter optimisation) on a subset of 1365 lightly regulated basins across CONUS and regionalization (i.e., parameter transfer) to the wider set of basins where either no streamflow observations are available or streamflow is more strongly impacted by water management (Cosgrove et al., 2024). The model was calibrated for the period 2016-10-01 to 2021-09-30 (NOAA, personal communication, 2025). In contrast to the setup used for forecasting, retrospective runs do not include data assimilation.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

For computational efficiency, we aggregated the hourly retrospective simulations to daily average values. This is not uncommon (e.g., Johnson et al., 2023; Towler et al., 2023), though we note the model runs operationally at an hourly timestep and is most commonly used to predict flood peaks in basins with a response time well below 24 hours. The model skill in simulating diurnal patterns will thus not be visible nor assessed in this study. Moreover, the goal of this work is to demonstrate the use of benchmarks in model evaluation, and the average daily NWM simulations provides a useful test case to do so.

2.2 Forcing data and streamflow observations

Though NWMv3.0 simulations are available without a need to run the model, we need certain meteorological data for the benchmarks used in this work (benchmarks are explained in Section 2.4). The Analysis of Record for Calibration (AORC) is an hourly ∼800-m-resolution gridded meteorological forcing dataset used as input to NWM retrospective simulations (Fall et al., 2023; Cosgrove et al., 2024), and thus used as input for the benchmarks in this work. We first aggregated the hourly gridded precipitation and 2-m air temperature to hourly basin averages using the areal mean. Precipitation was then aggregated from hourly to daily by summing the hourly amounts for each day from 1979-02-01 to 2023-02-01. For 2-m air temperature, we computed the daily mean. Streamflow observations from 1980-01-01 to 2023-12-31 were collected for approximately 5,000 GAGES-II gauges for which streamflow simulations are available (i.e., stream reaches are represented and gauges are active for the full simulation period) in the NWMv3.0 retrospective dataset (U.S. Geological Survey, 2025).

2.3 Model performance quantification

The Kling-Gupta efficiency (KGE; Gupta et al., 2009) was used to calibrate the NWMv3.0 on hourly timesteps (NOAA, personal communication, 2025):

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{1}$$

$$\alpha = \frac{\sigma_s}{\sigma_o}, \qquad \beta = \frac{\mu_s}{\mu_o}, \tag{2}$$

where r is the Pearson correlation coefficient and subscripts $_o$ and $_s$ indicate observations and simulations, respectively. To stay as close to the NWM setup as possible, we use the Kling-Gupta efficiency to quantify model performance in the remainder of this paper (though again note that we perform our analysis at daily time steps whereas the NWM was calibrated at hourly resolution). We repeated our analysis with the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970, presented in the Supporting Information) to investigate if our conclusions hold for a different metric.

2.4 Benchmarks

We compare the performance of the NWM to the performance of various simple benchmark models. The benchmark models are generated using the Python package `HydroBM` (Knoben, 2024) and cover various levels of complexity. At the simplest

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

end, these benchmarks are simple statistics calculated from the streamflow observations, which are then used as a predictor of streamflow on all time steps. One example is the long-term mean flow which, if used as a predictor of flow, returns a time series of constant values. Of intermediate complexity are benchmarks that rely on first calculating an average rainfall-runoff ratio and then applying this ratio to incoming precipitation to estimate the resulting runoff at each time step. The most complex benchmarks are still rather simple one- and two-parameter models whose parameters are optimized using a brute-force approach. A more in-depth explanation of the 17 different benchmark models used in this work can be found in Table S1.

Similar to Knoben (2024), 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" might be a more accurate description for most benchmarks, because no calibration takes place for most of them) and then used to predict the streamflow in an independent validation 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. The `HydroBM` package also includes a simple degree-day-based snow accumulation and melt routine, which we used with default parameters in snow-dominated basins.

2.5 Sampling uncertainty

Sampling uncertainty can be quantified with bootstrapping methods as implemented in the `gumboot` R package (Clark et al., 2021; Clark and Shook, 2021). The `gumboot` package works by creating a collection of synthetic hydrographs and calculating the score(s) of interest (such as KGE) from the observations and each synthetic hydrograph. We ran `gumboot` with the default settings as given in Clark et al. (2021). Briefly, this means that `gumboot` creates each synthetic hydrograph by dividing the period of record into water years (using October as the starting month and enforcing a minimum of 100 valid values within each water year) and sampling water years with replacement until the record length is reached. Using water years ensures that each sampled period is hydrologically independent, and the synthetic records are thus plausible hydrographs for the basin. With default settings `gumboot` returns 1000 synthetic hydrographs and associated NSE and KGE scores. We then define the sampling uncertainty as the difference between the 5^{th} and 95^{th} percentile of these scores.

We calculate the sampling uncertainty for each basin, for both the NWM simulations and each of the 17 benchmarks. This allows us to report both KGE scores and their associated uncertainty, and from this derive whether the accuracy of NWM simulations can be considered statistically different from the accuracy of the benchmarks. We report those results as Cumulative Distribution Functions (CDFs) that show that scores and uncertainty across the sample. We also report these results on a per-basin basis for the NWM and the best-performing benchmark. In this case, we use the Jaccard index (also known as the ratio of verification, critical success index, and Tanimoto index) to quantify the relative overlap of both uncertainty intervals. Assuming two uncertainty intervals, I_1 and I_2, defined as the difference between the 5^{th} (I^{p05}) to 95^{th} (I^{p95}) percentile estimates of KGE scores for the NWM (I_1) and benchmark (I_2):

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

$$\mathrm{J}(I_1, I_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|} = \frac{\text{overlap}}{\text{span}}, \tag{3}$$

$$\text{overlap} = \max\{0,\ \min(I_1^{p95}, I_2^{p95}) - \max(I_1^{p05}, I_2^{p05})\},$$

$$\text{span} = \max(I_1^{p95}, I_2^{p95}) - \min(I_1^{p05}, I_2^{p05}).$$

When $overlap$ = $span$, both sampling uncertainty intervals exactly overlap, and the performance of the NWM can be considered indistinguishable from the performance of the benchmark. When $overlap$ = 0, the uncertainty intervals do not overlap, and the performance of the NWM and benchmark simulations can thus be considered to be clearly different. We then need to further distinguish whether the NWM performance can be considered higher or lower than that of the benchmarks. Here we make the simplifying assumption that the 50^{th} percentile score estimate can be used to determine the relative positions of both uncertainty intervals. If the 50^{th} percentile estimate of NWM performance is higher than the 50^{th} percentile estimate of benchmark performance, we consider the NWM to perform better than the benchmark (and vice versa). How much better (or worse) the performance of the NWM is, can then be quantified using Eq. 1. High values of J indicate a large amount of overlap (with complete overlap at J = 1) between the two distributions (i.e., smaller distinguishable differences), whereas low values of J indicate a small amount of overlap and clearer differences between the two distributions (no overlap at J = 0). A schematic overview of the methodology can be found in Fig. 1.

3 Results

3.1 Aggregated performance

Figure 2a and 2b show 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. First, for both the calibration and evaluation period, the NWM (black line) reaches higher KGE scores considerably more often than any of the benchmarks (colored lines). However, NWM performance also shows a tendency to decline quickly at lower KGE values, suggesting that there are locations where NWM performance is not as high as that of some of the benchmarks. Second, three benchmarks of note are BM01 (for performing quite poorly), and benchmarks BM07 and BM17 (for performing rather well).

BM01 (the mean flow benchmark) can be found as a nearly vertical line at KGE $= 1 - \sqrt{(2)} \approx -0.41$ (dark blue in the top row, see also Fig. 2c). This is the traditional choice of benchmark model, derived from the original formulation of the Nash-Sutcliffe efficiency, and it is the only benchmark that shows no spatial variability at all during calibration (there is some variability during evaluation, because the mean flow calculated from the calibration data is not always close to the actual mean flow during evaluation). Comparison of this CDF to all others highlights the point made by Schaefli and Gupta (2007): the mean flow is not an equally hard-to-beat benchmark in all basins, and location-specific benchmarks are needed to set more locally appropriate expectations for models (see also Knoben, 2024).

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

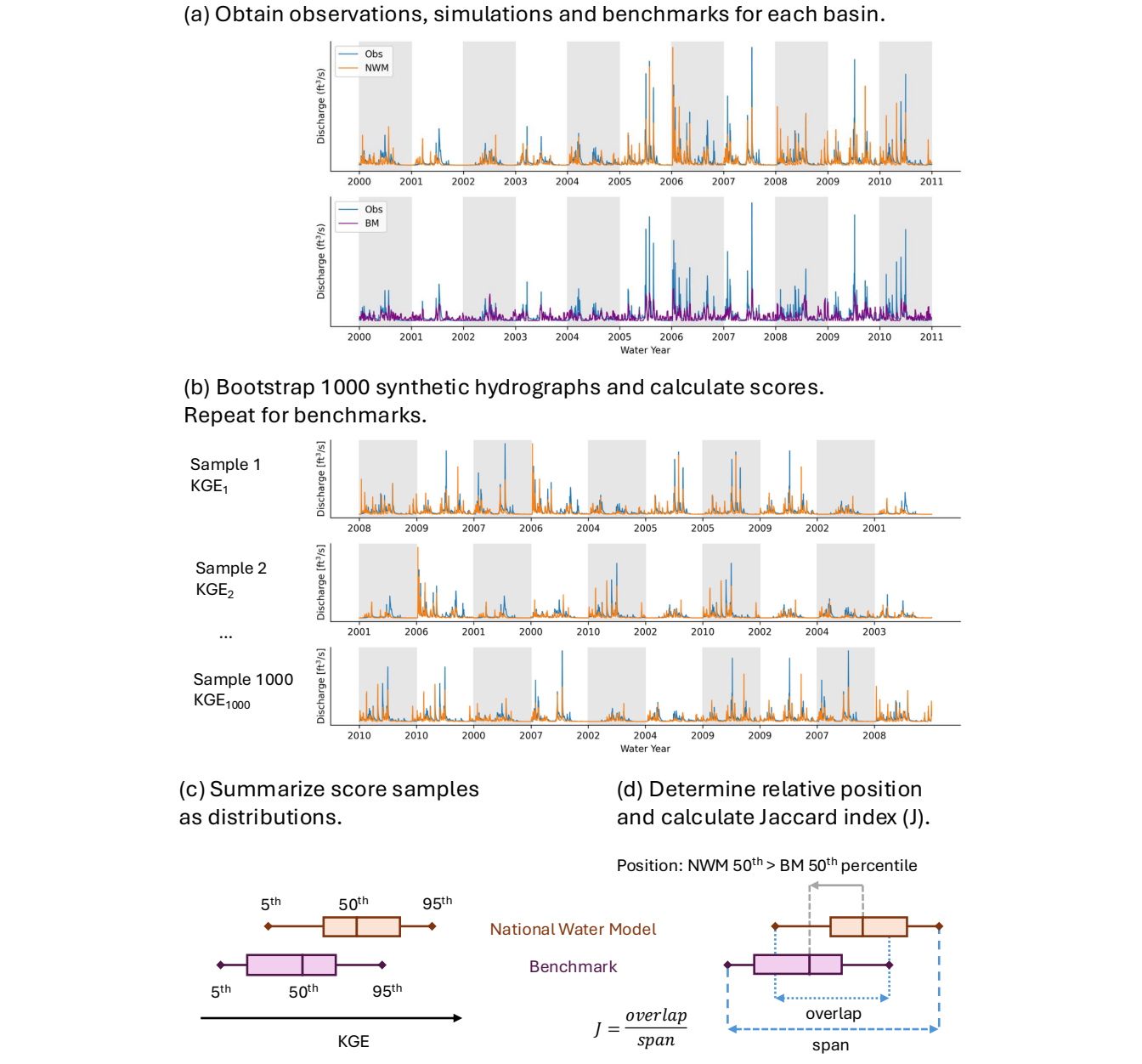

Figure 1. Schematic overview of methodology. (a) Example selection of water years, showing observations as well as NWM simulations (top) and one of the benchmark simulations (bottom) for an arbitrary gauge (USGS ID 01037380). Water years indicated with alternating grey/white blocks. (b) Examples of synthetic hydrographs obtained from sampling water years with replacement. Water years indicated with alternating grey/white blocks. (c) Schematic representation of the 1000 KGE samples for the NWM and the benchmark, summarized as boxplots. (d) Overview of the terminology and method used to quantify relative overlap of the NWM and benchmark KGE samples.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

BM07 (the daily mean flow benchmark in red; see also Fig. 2d) is computed by taking the mean flow on each Julian day in the calibration period and appending these values to create a year-long timeseries, which is then repeated for each year of the full simulation period. While its CDF does not cover scores as high as the NWM CDF, this benchmark equally does not lead to KGE scores that are as low as some of those obtained by the NWM: during calibration, the NWM CDF covers a range of (roughly) <-5, 1], whereas the CDF of BM07 covers a more restricted range of (roughly) [0,1].

For unseen data (evaluation) the BM07 CDF does not stand out compared to the other benchmarks, possibly due to the somewhat limited amount of data (maximum 5 years) used to compute the benchmark. BM17 (the adjusted smoothed precipitation benchmark in light blue; see also Fig. 2e) is a simple 2-parameter model that aims to capture three dominant facets of catchment functioning: partitioning of incoming precipitation into streamflow and sink terms, as well as time delay and attenuation of the resulting runoff (Schaefli and Gupta, 2007). Its CDF is quite similar to that of the NWM but more constrained; the KGE values for this benchmark are neither as high nor as low as those obtained by the NWM. Combined, this suggests that there are basins where the NWM performance is constrained in some way that the benchmarks are not.

Figure 3 shows the sampling uncertainty associated with the benchmarks and NWM simulations using data from the evaluation period. To save space, a number of benchmarks have been omitted: BM01 and BM02 (mean and median flow) as well as BM10 (rainfall-runoff ratio to annual) have, in the majority of cases, limited performance and little can be learned from these; BM03 and BM04 (annual mean and median flow) use annual flow statistics as a predictor and can by definition not be used for unseen data; BM09 (rainfall-runoff ratio applied to all timesteps) is conceptually very similar to BM01 and has been omitted for the same reason.

As shown in earlier work (Clark et al., 2021), the sampling uncertainty around KGE scores can be substantial. In the case of the NWM (black line with grey uncertainty bounds) there is a broad inverse correlation between the KGE score and associated uncertainty bounds, though considerable scatter is present. This emphasizes the strong need to evaluate models while accounting for sampling uncertainty. In numerous basins, the KGE scores obtained by the NWM are strongly conditional on the idiosyncrasies of the evaluation period, and the same model instantiation might be evaluated quite differently if a different time period were to be used. The benchmarks show varying levels of sampling uncertainty. Some are mostly insensitive to data selection (e.g., BM06, BM08), whereas others are either highly sensitive (e.g., BM12, BM16), mostly robust but occasionally sensitive (e.g., BM06, BM08), or somewhere in between (e.g., BM07, BM17). The CDFs and uncertainty bounds should not be directly compared between the different subplots, but a general idea of the widths of these uncertainty intervals is helpful for understanding the results in the next section.

3.2 Spatial patterns

While CDFs of performance scores can be helpful to quickly compare performance differences across the full sample of basins, such approaches do not facilitate a basin-by-basin comparison of differences. Figures 4a and 4d therefore show a spatial overview of model and benchmark performance during the evaluation period, using `gumboot`'s estimated 50^{th} percentile KGE score for both. For simplicity, we only assess the evaluation performance of the best benchmark in each basin (in other words, Figure 4b is a composite of different benchmarks selected for having the highest 50^{th} percentile KGE score). Both maps

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

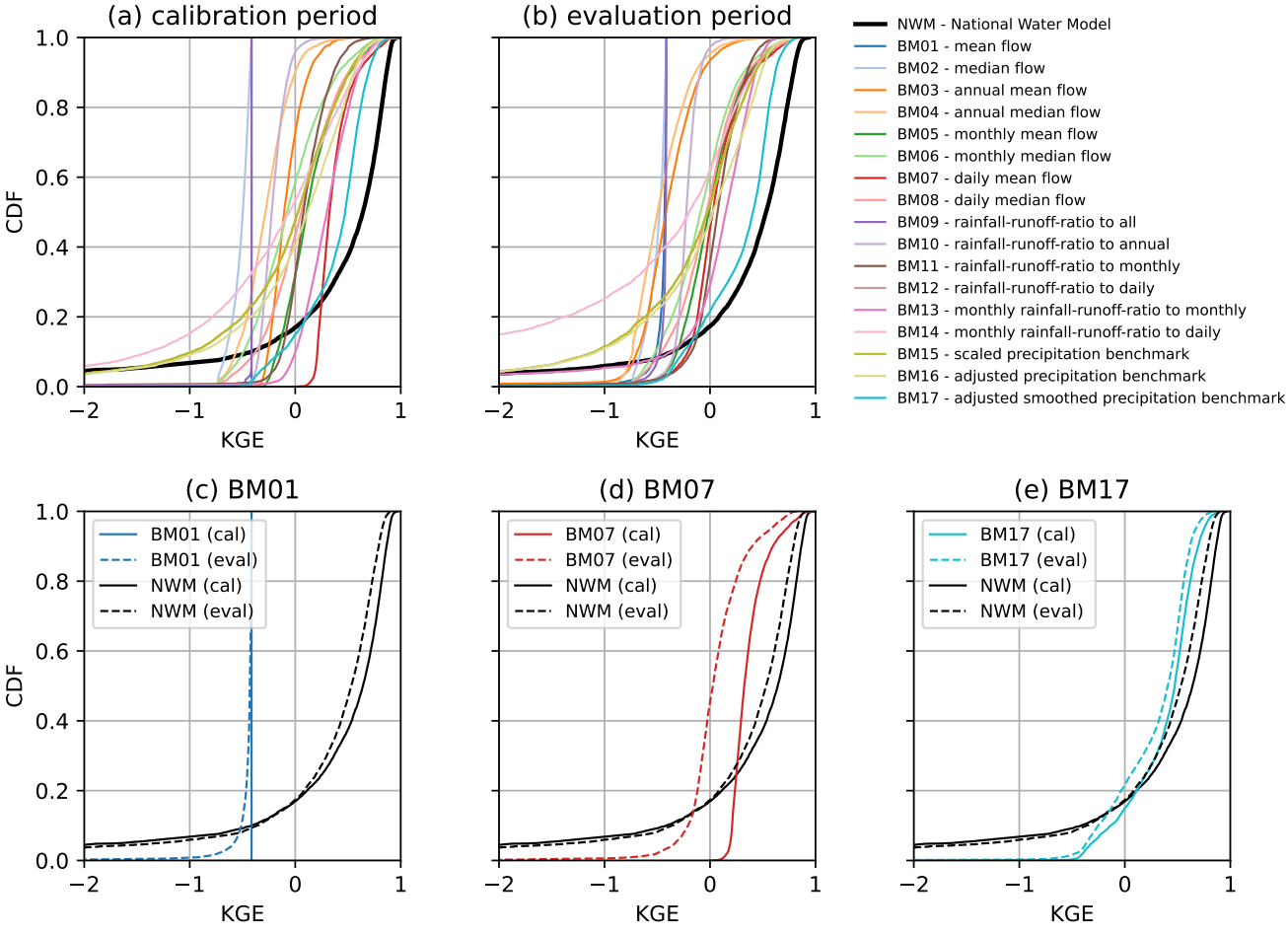

Figure 2. Cumulative Distribution Function (CDF) plot of the Kling–Gupta Efficiency (KGE) scores for the NWMv3.0 and 17 simple benchmarks during (a) calibration and (b) evaluation. (c-e) Individual plots of the three benchmarks discussed in the text.

confirm the broad statement suggested by the various CDFs, namely that the NWM spans a wider range of performance scores than the benchmarks. This suggests that in a considerable number of places the NWM outperforms a fairly taxing collection of benchmarks, even when sampling uncertainty is accounted for. However, it also suggests that in numerous places the NWM does not replicate certain aspects of the flow regime to the extent that the benchmarks do.

Figures 4b, 4c, 4e and 4f show the relative overlap of the sampling uncertainty intervals of the NWM and best benchmark. Overlap is quantified with the Jaccard index (Eq. 3) and separated into cases where the estimated 50^{th} percentile KGE score of the NWM is higher than that of the best benchmark (4b, 4c) and vice versa (4e, 4f). These results are separated into basins used for calibration of the NWM parameters (4b, 4e), and cases where NWM parameters were regionalized (4c, 4f). For both sets of plots, the colored stations are complementary: a station plotted in green in Figure 4b (or Fig. 4c) will appear as a yellow

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

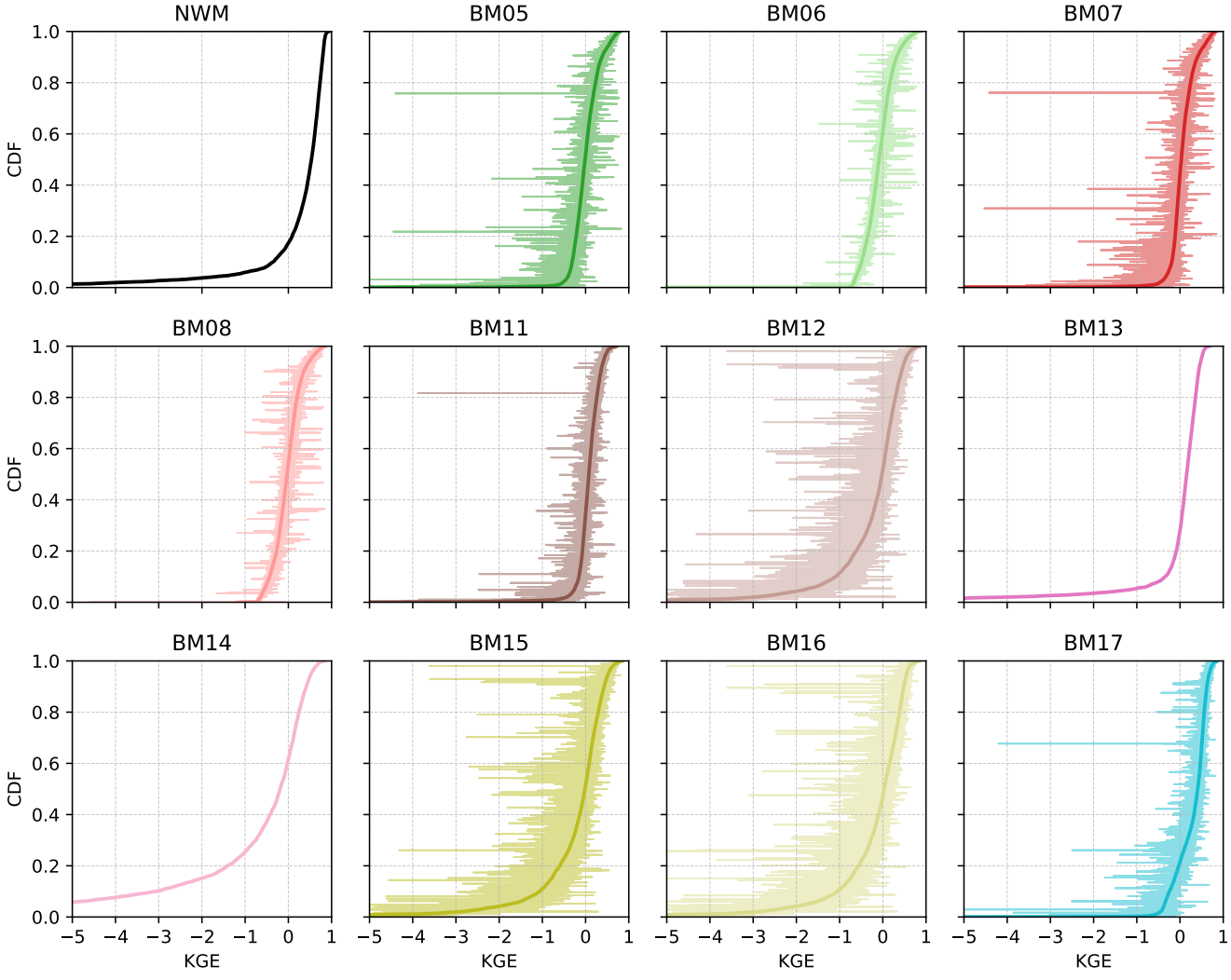

Figure 3. Cumulative Distribution Function (CDF) plot of the Kling–Gupta Efficiency (KGE) scores of the evaluation period. The NWMv3.0 KGE scores are in black, and the KGE scores for the simple benchmarks in colors. Sampling uncertainty (defined as the difference between the 5^{th} and 95^{th} percentile KGE estimate) in lighter colors. See Fig. 2 for a description of the benchmarks.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

dot in Figure 4e (or Fig. 4f) and vice versa. Note that no overlap (Jaccard index = 0; dark green and bright red) indicates that the distributions of KGE scores are clearly separate (in other words, the NWM score is either clearly higher or lower than the benchmark score), whereas lighter colors indicate that the performance of the NWM and benchmark are closer together.

Figure 4b shows that in approximately 70% of calibration basins the NWM outperforms the benchmarks. In most basins this is a clear improvement (Jaccard index ≈ 0). Basins where the KGE distributions of the NWM and best benchmark partly overlap are mostly found in the central mountainous and drier regions. Figure 4e shows the remaining 30% of calibration basins where the benchmarks outperform the NWM. Here too the overlap between the KGE distributions is mostly low, which shows that in these basins the benchmarks tend to obtain clearly higher scores than the NWM. Clusters of basins where the benchmarks outperform the NWM seems mostly concentrated in the interior west (broadly inland of the western coastal mountain ranges until somewhat east of the 100^{th} meridian) and the Appalachian Piedmont, with scattered occurrences elsewhere.

These patterns are reinforced in Figures 4c and 4f, showing the performance of the NWM in basins where its parameters were regionalized (i.e., not calibrated). The NWM mainly outperforms the benchmarks along the western coast and in the humid eastern part of the US. In contrast, the benchmarks perform better in the interior west and the Appalachians, with the appearance of a new cluster of strong performance in central Florida and an increase in scattered basins. Notably, the benchmarks outperform the NWM in almost half of the regionalization basins, with clear regional patterns.

As shown in the Supportining Information, these findings generally hold when the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) is used to quantify model and benchmark performance, but with a few important caveats. First, the benchmarks show a tendency towards lower NSE scores and their CDFs are further away from the NWM CDF (Figures S1, S2). Second, the NWM outperforms the benchmarks in more basins when NSE is used to quantify model performance (the NWM is better in 79.3% of calibration basins and in 63.4% of regionalization basins; Figure S3). This is somewhat surprising, given that the benchmarks are identical in both cases and the NWM was not calibrated on NSE, and points to a need for further work on robust model evaluation practices. Preliminary analysis suggests that these differences are driven by the different sensitivities of NSE and KGE to the bias, variability and correlation components (see e.g., Gupta et al., 2009; Knoben et al., 2019; Lamontagne et al., 2020). In at least some basins, the benchmarks perform clearly better on bias and much worse on correlation than the NWM, and because correlation errors are weighted more heavily in NSE, this results in a larger difference in NSE scores than in KGE scores.

4 Discussion

We demonstrated how simple benchmarks can be used to assess the performance of large-domain hydrologic models. As our test case, we compared the NWMv3.0 daily-averaged retrospective simulation against the performance of 17 simple benchmark models across approximately 5000 basins in the United States. In basins used for model calibration, the benchmarks outperform the NWM in approximately 30% of basins. The benchmarks perform primarily better in the interior mountainous and drier plains areas in the west as well as in the Appalachians. This pattern, with the addition of a cluster of basins in central Florida, appears even clearer in basins where the NWM parameters were regionalized, and the benchmarks outperform the NWM in

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

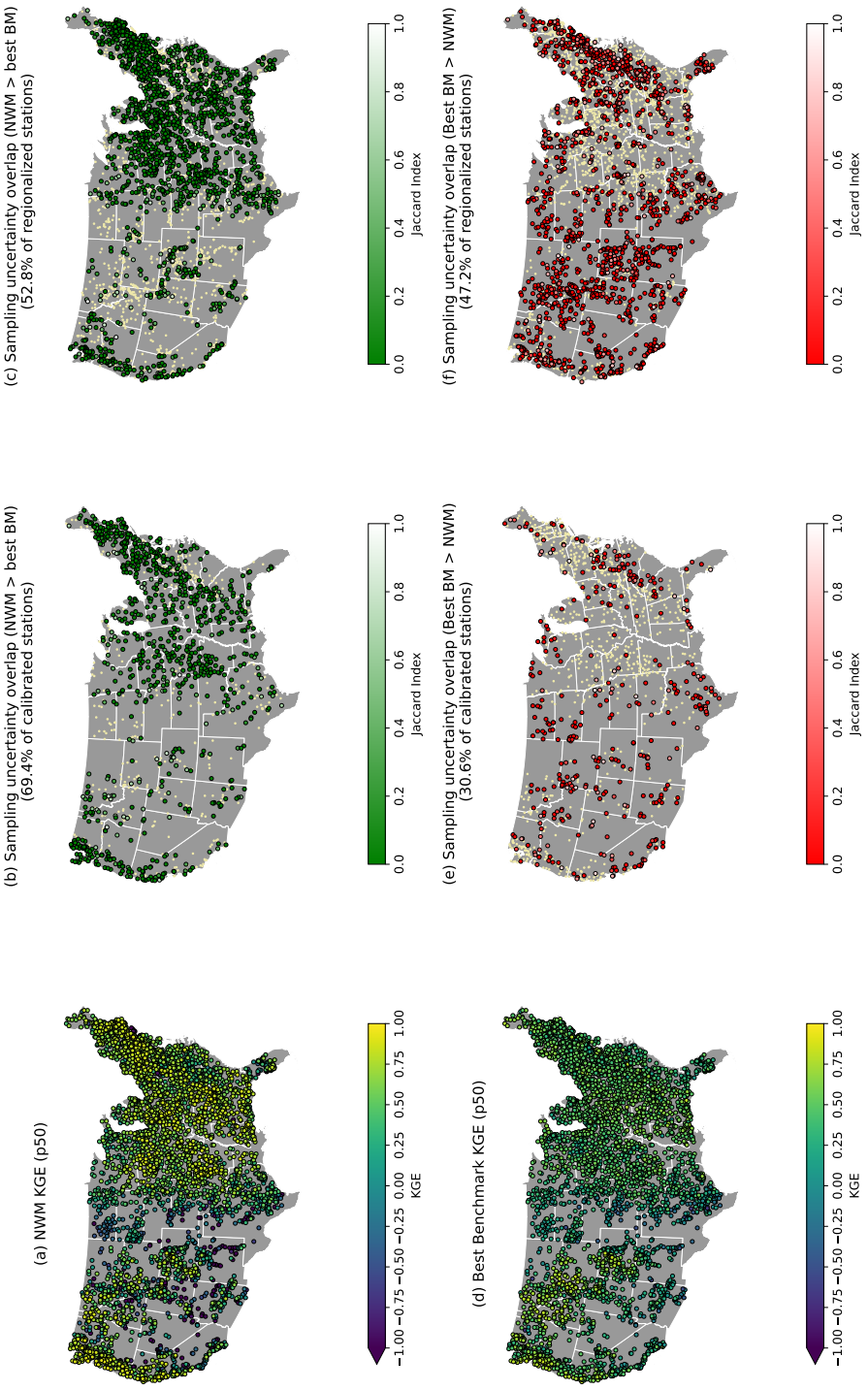

Figure 4. Overview of spatial patterns in model and benchmark performance, and overlap. (a,d) Estimated 50th percentile KGE score for NWM and best benchmark respectively. (b,e) Jaccard index showing overlap between sampling uncertainty intervals where the 50th percentile KGE score for NWM >benchmark, and NWM <benchmark, respectively, for gauges used for model calibration. (c,f) Jaccard index showing overlap between sampling uncertainty intervals where the 50th percentile KGE score for NWM >benchmark, and NWM <benchmark, respectively, for gauges used for model regionalization. Borders obtained from Commission for Environmental Cooperation (CEC) (2022).

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

almost 50% of the basins. These patterns are different from where KGE scores suggest that the model does poorly (Fig. 4a). Based on KGE scores alone, one might conclude that the model does worst in the drier southwestern and central areas, but when performance is compared against benchmarks, more regions stand out as areas where improvements may be possible.

These results are broadly consistent with various evaluations of earlier versions of the NWM. For example, Towler et al. (2023) find that on daily time steps the NWMv2.1 outperforms a single benchmark in 80% of cases, and that NWM performance is better in natural basins than regulated ones (i.e., basins where parameters are regionalized, also shown by Abdelkader et al., 2023, though at hourly time steps). In ecological terms, one of the regions where the benchmarks provide better simulations than the NWMv3.0 broadly coincides with the Mediterranean California, North American Desserts, Temperate Sierras, and Great Plains eco-regions (Commission for Environmental Cooperation, 1997). This aligns with results from Johnson et al. (2023), who found that the performance of the NWMv2.0 can be improved in drier climates with predominantly low vegetation.

While more in depth study is needed to understand the contributing factors, the nature of the benchmarks lets us speculate about potential improvements to the modeling chain. Three main lines of investigation may be worthwhile, focusing on model inputs, model structure, and model parametrization/regionalization. Large-domain parameter estimation has long been an open challenge, but existing (e.g., Samaniego et al., 2010) and promising recent advances (e.g., Shen et al., 2023; Tang et al., 2025; Farahani et al., 2025) have not yet been implemented in most large-domain modeling chains. Regionalization of model parameters is similarly challenging (e.g., Merz and Blöschl, 2004; Pool et al., 2021; Yang et al., 2023). The relative success of the benchmarks during both calibration (effective in 30% of basins) and regionalization (effective in 50% of basins) may suggest that improvements to parameter optimization and regionalization are possible.

The strong regional patterns in where the benchmarks outperform the models suggest solutions may need to be found more locally as well. For example, in the current NWM setup, parameters are regionalized for regulated basins. The NWM currently accounts for the location of more than 5000 reservoirs but does not include any operating rules for these reservoirs. Instead, data assimilation is used to correct and align model states with observations during forecasting for several hundred of these reservoirs (Cosgrove et al., 2024). The relative success of the benchmarks in the regulated basins suggests that some aspects of the resulting regulated streamflow are relatively predictable and that implementing a rudimentary reservoir operations module may be possible. Similarly, the relative success of the benchmarks in drier regions may point to a need to account for dry-region processes such as channel infiltration and transmission losses. Improvements to the representation of shallow aquifer systems (e.g., in the Northern Appalachian Mountains and Appalachian Piedmont; Rutledge and Mesko, 1996; Swain et al., 2004), low-lying coastal areas and wetlands (e.g., central Florida), and snow pack dynamics (e.g., the western mountains), and surface depression storage (e.g., the prairie pothole region in North Dakota, South Dakota, Minnesota and Iowa) might also be needed.

However, the relative success of the benchmarks in these regions may also point to potential issues with the forcing data (see e.g. Quansah et al., 2025, who identify issues with convective summer precipitation in the NWMv2.1 forcing data over Alabama). The benchmarks are only minimally (or not at all) constrained by a need to respect mass and energy balances within the system, and will typically produce relatively unbiased simulations with larger variability and correlation errors (see Figures S8 and S9). The model instead is bound by a need to partition its precipitation input correctly between storage, streamflow and evaporation, and may thus be more vulnerable to biases in the forcing data (compare with Cosgrove et al., 2024, who show that

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

the NWMv2.1 has considerable bias in its simulations). Regions where the benchmarks outperform the model may thus also be locations where biases in the forcing data limit the model's ability to produce accurate streamflow simulations.

The type of benchmark may give some hints about the kind of problem the model encounters in a given region. Preliminary analysis (Figures S4-S7) suggests that there are spatial patterns in the type of benchmark that provides the highest accuracy in each region. Streamflow-based benchmarks (Group 1) dominate in the Rocky Mountains, suggesting that the streamflow regimes here are relatively stable year-to-year. Runoff-ratio benchmarks (Group 2) are often the best benchmark in the drier parts of the western CONUS, suggesting that the partitioning of precipitation into streamflow and other components is relatively predictable in these basins, but modulated by the amount of incoming precipitation. The last group of benchmarks (very simple models) are often the most accurate benchmark in the wetter parts of the western CONUS as well as in the east. However, local analysis and comparison of model simulations against the benchmarks remains needed in order to understand which components of the simulations are better captured by the benchmarks, and what this means for potential improvements to the modeling chain. Particularly with the recent increase in large-sample studies, where results are predominantly shown as maps of performance scores and associated Cumulative Distribution Functions, there is a risk that the performance scores become a goal in themselves while locally poor model performance goes undetected. Benchmarks provide a convenient way of quickly identifying areas where improvements may be possible and, critically, these are not always the same regions where we find lower model performance scores.

5 Conclusions

We used an ensemble of simple benchmarks to provide context for the performance of a large-domain water model. We also account for sampling uncertainty in this work, but results suggest that in most basins the differences in performance between the National Water Model v3.0 and the benchmarks are large enough that this is only a minor concern. However, sampling uncertainty remains important in cases where models perform similarly. The benchmarks suggest that there are considerable constraints on the model's performance in approximately one-third of the basins used for model calibration and in approximately half of the basins where model parameters are regionalized. The areas where the benchmarks outperform the model only partially overlap with areas where the model achieves lower KGE scores, and this suggests that improvements may be possible in more regions than a first glance at model performance values may indicate. In cases where the benchmarks outperform the model, the nature of the benchmarks may suggest which elements of the modeling chain could be improved but it remains difficult to go beyond listing broad hypotheses. In-depth model evaluation thus remains necessary to identify which aspects of the simulations the benchmarks simulate more accurately than the model does, and what this implies for potential ways to improve the model. A key advantage of using these benchmarks is that they are straightforward and fast to compute, particularly compared to the cost of configuring and running the model. This makes benchmarking a valuable tool that can complement more detailed model evaluation techniques by quickly identifying areas that should be investigated more thoroughly.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

Code and data availability. Streamflow observations were obtained on Mar 31, 2025 from the United States Geological Survey U.S. Geological Survey (2025). The NOAA National Water Model CONUS Retrospective Dataset was accessed on May 28, 2024 (AORC forcing) and Aug 31, 2024 (NWMv3.0 simulations) from https://registry.opendata.aws/nwm-archive. The benchmarks were calculated using the Python package `HydroBM` (Knoben, 2024), and the sampling uncertainty with the R package `gumboot` (Clark et al., 2021; Clark and Shook, 2021). Intermediate results (CSV files containing the sampling uncertainty values for the National Water Model as well as the benchmarks) and code to create the figures in this manuscript and the Supporting Information are available on Zenodo (Gründemann et al., 2025).

Author contributions. **Gaby Gründemann:** Conceptualization, Methodology, Software, Data Curation, Writing - Review & Editing, Visualization. **Wouter Knoben:** Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Yalan Song:** Data Curation, Software, Writing - Review & Editing. **Katie van Werkhoven:** Conceptualization, Data Curation, Writing - Review & Editing. **Martyn Clark:** Conceptualization, Methodology, Supervision, Writing - Review & Editing, Project administration, Funding acquisition.

Competing interests. The authors declare there are no conflicts of interest for this manuscript.

Acknowledgements. This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

References

Abdelkader, M., Temimi, M., and Ouarda, T. B.: Assessing the National Water Model's Streamflow Estimates Using a Multi-Decade Retrospective Dataset across the Contiguous United States, Water, 15, 2319, https://doi.org/10.3390/w15132319, 2023.

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrology and Earth System Sciences, 24, 535–559, https://doi.org/10.5194/hess-24-535-2020, 2020.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., Van Den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, Journal of Hydrometeorology, 16, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

Beven, K.: Benchmarking hydrological models for an uncertain future, Hydrological Processes, 37, e14 882, https://doi.org/https://doi.org/10.1002/hyp.14882, 2023.

Clark, M. P. and Shook, K.: gumboot: Bootstrap Analyses of Sampling Uncertainty in Goodness-of-Fit Statistics, https://github.com/CH-Earth/gumboot, R package version 1.0.1 (accessed 2024-09-04), 2021.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, https://doi.org/https://doi.org/10.1029/2007WR006735, 2008.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/https://doi.org/10.1029/2020WR029001, e2020WR029001 2020WR029001, 2021.

Commission for Environmental Cooperation: Ecological Regions of North America: Toward a Common Perspective, ISBN 2-922305-18-X, http://www.cec.org/files/documents/publications/1701-ecological-regions-north-america-toward-common-perspective-en.pdf, 1997.

Commission for Environmental Cooperation (CEC): North American Atlas – Political Boundaries, http://www.cec.org/north-american-environmental-atlas/political-boundaries-2021/, statistics Canada, United States Census Bureau, Instituto Nacional de Estadística y Geografía (INEGI), 2022.

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., Clark, E., Cabell, R., Casiday, N., Cui, Z., et al.: NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling, JAWRA Journal of the American Water Resources Association, 60, 247–272, 2024.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., and Kusche, J.: Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin, Hydrology and Earth System Sciences, 28, 2259–2295, https://doi.org/10.5194/hess-28-2259-2024, 2024.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, https://doi.org/10.1080/02626660903526292, 2010.

Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., Trypaluk, C., Wu, W., and Miller, D.: The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation, JAWRA Journal of the American Water Resources Association, 59, 1246–1272, 2023.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

Farahani, M. A., Wood, A. W., Tang, G., and Mizukami, N.: Calibrating a large-domain land/hydrology process model in the age of AI: the SUMMA CAMELS emulator experiments, Hydrology and Earth System Sciences, 29, 4515–4537, https://doi.org/10.5194/hess-29-4515-2025, 2025.

Gründemann, G., Knoben, W., Song, Y., van Werkhoven, K., and Clark, M.: Data for "Separating Signal from Noise in Large- Domain Hydrologic Model Evaluation: Benchmarking model performance under sampling uncertainty", https://doi.org/10.5281/zenodo.18028488, 2025.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations : elements of a diagnostic approach to model evaluation, Hydrological Processes, 3813, 3802–3813, https://doi.org/https://doi.org/10.1002/hyp.6989, 2008.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, iSBN: 0022-1694 Publisher: Elsevier B.V., 2009.

Gupta, H. V., Clark, M. P., Vrugt, J. a., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, https://doi.org/10.1029/2011WR011044, 2012.

Harrigan, S., Zsoter, E., Cloke, H., Salamon, P., and Prudhomme, C.: Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System, Hydrology and Earth System Sciences, 27, 1–19, https://doi.org/10.5194/hess-27-1-2023, 2023.

Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., Kindl Da Cunha, L., Jennings, K. S., Clarke, K. C., Mazrooei, A., and Yeghiazarian, L.: Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous Formulations and Diagnostic Model Selection, Journal of Geophysical Research: Atmospheres, 128, e2023JD038 534, https://doi.org/10.1029/2023JD038534, 2023.

Klotz, D., Gauch, M., Kratzert, F., Nearing, G., and Zscheischler, J.: Technical Note: The divide and measure nonconformity – how metrics can mislead when we evaluate on different data partitions, Hydrology and Earth System Sciences, 28, 3665–3673, https://doi.org/10.5194/hess-28-3665-2024, 2024.

Knoben, W. J. M.: Setting expectations for hydrologic model performance with an ensemble of simple benchmarks, Hydrological Processes, 38, e15 288, https://doi.org/https://doi.org/10.1002/hyp.15288, 2024.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Knoben, W. J. M., Raman, A., Gründemann, G. J., Kumar, M., Pietroniro, A., Shen, C., Song, Y., Thébault, C., Van Werkhoven, K., Wood, A. W., and Clark, M. P.: Technical note: How many models do we need to simulate hydrologic processes across large geographical domains?, Hydrology and Earth System Sciences, 29, 2361–2375, https://doi.org/10.5194/hess-29-2361-2025, 2025.

Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, Water Resources Research, 48, https://doi.org/10.1029/2011WR011534, 2012.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11 344–11 354, https://doi.org/10.1029/2019WR026065, 2019.

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data, Water Resources Research, 56, e2020WR027 101, https://doi.org/https://doi.org/10.1029/2020WR027101, e2020WR027101 2020WR027101, 2020.

Legates, D. R. and Mccabe, G. J.: A refined index of model performance: A rejoinder, International Journal of Climatology, 33, 1053–1056, https://doi.org/10.1002/joc.3487, 2013.

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

McCuen, R. H., Knight, Z., and Cutter, A. G.: Evaluation of the Nash–Sutcliffe Efficiency Index, Journal of Hydrologic Engineering, 11, 597–602, https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597), 2006.

Merz, R. and Blöschl, G.: Regionalisation of catchment model parameters, Journal of Hydrology, 287, 95–123, https://doi.org/10.1016/j.jhydrol.2003.09.028, 2004.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

NOAA: The National Water Model, https://water.noaa.gov/about/nwm, 2025.

Pool, S., Vis, M., and Seibert, J.: Regionalization for Ungauged Catchments — Lessons Learned From a Comparative Large-Sample Study, Water Resources Research, 57, e2021WR030 437, https://doi.org/10.1029/2021WR030437, 2021.

Quansah, J., Doria, R., and Fall, S.: Evaluating the Performance of the National Water Model: A Spatiotemporal Analysis of Streamflow Forecasting, Water, 17, 2950, https://doi.org/10.3390/w17202950, 2025.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resources Research, 52, 7779–7792, https://doi.org/10.1002/2016WR019430, 2016.

Ritter, A. and Muñoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, Journal of Hydrology, 480, 33–45, https://doi.org/10.1016/j.jhydrol.2012.12.004, publisher: Elsevier B.V., 2013.

Rutledge, A. T. and Mesko, T. O.: Estimated hydrologic characteristics of shallow aquifer systems in the Valley and Ridge, the Blue Ridge, and the Piedmont Physiographic Provinces based on analysis of streamflow recession and base flow, Professional Paper 1422-B, United States Geological Survey, https://doi.org/https://doi.org/10.3133/pp1422B, 1996.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 1–25, https://doi.org/10.1029/2008WR007327, 2010.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, https://doi.org/https://doi.org/10.1002/hyp.6825, 2007.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrological Processes, 15, 1063–1064, https://doi.org/https://doi.org/10.1002/hyp.446, 2001.

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H.: Upper and lower benchmarks in hydrological modelling, Hydrological Processes, 32, 1120–1125, https://doi.org/https://doi.org/10.1002/hyp.11476, 2018.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify

https://doi.org/10.5194/egusphere-2025-6460
Preprint. Discussion started: 2 February 2026

machine learning and physical models for geosciences, Nature Reviews Earth & Environment, 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Song, Y., Bindas, T., Shen, C., Ji, H., Knoben, W. J. M., Lonzarich, L., Clark, M. P., Liu, J., Van Werkhoven, K., Lamont, S., Denno, M., Pan, M., Yang, Y., Rapp, J., Kumar, M., Rahmani, F., Thébault, C., Adkins, R., Halgren, J., Patel, T., Patel, A., Sawadekar, K. A., and Lawson, K.: High-Resolution National-Scale Water Modeling Is Enhanced by Multiscale Differentiable Physics-Informed Machine Learning, Water Resources Research, 61, e2024WR038 928, https://doi.org/10.1029/2024WR038928, 2025.

Swain, L. A., Mesko, T. O., and Hollyday, E. F.: Summary of the hydrogeology of the Valley and Ridge, Blue Ridge, and Piedmont Physiographic Provinces in the eastern United States, Professional Paper 1422-A, United States Geological Survey, https://doi.org/https://doi.org/10.3133/pp1422A, 2004.

Tang, G., Wood, A. W., and Swenson, S.: On Using AI-Based Large-Sample Emulators for Land/Hydrology Model Calibration and Regionalization, Water Resources Research, 61, e2024WR039 525, https://doi.org/10.1029/2024WR039525, 2025.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrology and Earth System Sciences, 27, 1809–1825, https://hess.copernicus.org/articles/27/1809/2023/, 2023.

U.S. Geological Survey: U.S. Geological Survey National Water Information System Database, https://doi.org/10.5066/F7P55KJN, https://doi.org/10.5066/F7P55KJN, accessed March 21, 2025. Data download directly accessible at https://waterdata.usgs.gov/nwis/dv, 2025.

Van Jaarsveld, B., Wanders, N., Sutanudjaja, E. H., Hoch, J., Droppers, B., Janzing, J., Van Beek, R. L. P. H., and Bierkens, M. F. P.: A first attempt to model global hydrology at hyper-resolution, Earth System Dynamics, 16, 29–54, https://doi.org/10.5194/esd-16-29-2025, 2025.

Yang, X., Li, F., Qi, W., Zhang, M., Yu, C., and Xu, C.-Y.: Regionalization methods for PUB: a comprehensive review of progress after the PUB decade, Hydrology Research, 54, 885–900, https://doi.org/10.2166/nh.2023.027, 2023.