https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


# On the importance of observation uncertainty when evaluating and comparing models: a hydrological example

Jerom P.M. Aerts[1], Jannis M. Hoch[2,3], Gemma Coxon[4], Nick C. van de Giesen[1], and Rolf W. Hut[1]

[1]Department of Water Management, Civil Engineering and Geoscience, Delft University of Technology, Delft, the Netherlands
[2]Department of Physical Geography, Utrecht University, Utrecht, the Netherlands
[3]Fathom, Bristol, United Kingdom
[4]Geographical Sciences, University of Bristol, Bristol, United Kingdom

**Correspondence:** Jerom Aerts (j.p.m.aerts@tudelft.nl)

**Abstract.** The comparison of models in geosciences involves refining a single model or comparing various model structures. However, such model comparison studies are potentially invalid without considering the uncertainty estimates of observations in evaluating relative model performance. The temporal sampling of the observation and simulation time series is an additional source of uncertainty as a few observation and simulation pairs, in the form of outliers, might have a disproportionate effect on the model skill score. In this study we highlight the importance of including observation uncertainty and temporal sampling uncertainty when comparing or evaluating hydrological models.

In hydrology, large-sample hydrology datasets contain a collection of catchments with hydro-meteorological time series, catchment boundaries and catchment attributes that provide an excellent test-bed for model evaluation and comparison studies. In this study, two model experiments that cover different purposes for model evaluation are set up using 396 catchments from the CAMELS-GB dataset. The first experiment, *intra-model*, mimics a model refinement case by evaluating the streamflow estimates of the distributed wflow_sbm hydrological model with and without additional calibration. The second experiment, *inter-model*, is a model comparison based on the streamflow estimates of the distributed PCR-GLOBWB and wflow_sbm hydrological models.

The temporal sampling uncertainty, the result of outliers in observation and simulation pairs, is found to be substantial throughout the case study area. High temporal sampling uncertainty indicates that the model skill scores used to evaluate model performance are heavily influenced by only a few data points in the time series. This is the case for half of the simulations (210) of the first intra-model experiment and 53 catchment simulations of the second inter-model experiment as indicated by larger sampling uncertainty than the difference in the KGE-NP model skill score. These cases highlight the importance of reporting and determining the cause of temporal sampling uncertainty before drawing conclusions on large-sample hydrology based model performance. The streamflow observation uncertainty analysis shows similar results. One third of the catchments simulations (123) of the intra-model experiment contains smaller streamflow simulation differences between models than streamflow observation uncertainties, compared to only 4 catchment simulations of the inter-model experiment due to larger differences between streamflow simulations. These catchments simulations should be excluded before drawing conclusions based on large-samples of catchments. The results of this study demonstrate that it is crucial for benchmark efforts based on

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


large-samples of catchments to include streamflow observation uncertainty and temporal sampling uncertainty to obtain more robust results.

## 1 Introduction

Many fields in geoscience rely on uncertain data to accurately estimate states and fluxes that support decision-making. One challenging aspect of hydrological modelling in particular is the large spatial and temporal landscape and hydrological heterogeneity (e.g. Gao et al. (2018)). Capturing this large variety in landscape and hydrological heterogeneity when evaluating or comparing hydrological models can be achieved through the use of so called large-sample catchment hydrology datasets.

These large-sample datasets contain hydro-meteorological timeseries, catchment boundaries and catchment attributes for a large number of catchments. They are complemented with streamflow observations at the catchment outlet and meteorological forcing data such as precipitation and temperature. The datasets are created by applying a consistent methodology across all catchments. Recent large-sample datasets follow the structure introduced by Addor et al. (2017) in the form of the CAMELS(-US) dataset. A recent effort by Kratzert et al. (2022) combined all available national CAMELS datasets in the overarching CARAVAN dataset for global consistency and boosting accessibility through data access via Google Earth Engine.

The accessibility of large-sample data triggered a wealth of research as discussed in the overview by Addor et al. (2020), including as a test-bed for hydrological model evaluation and model comparison studies (e.g. Mizukami et al. (2017); Rakovec et al. (2019); Lane et al. (2019); Feng et al. (2022)). The benefits of using large-sample datasets are that by including large samples of catchments, the robustness of model results is tested (Andréassian et al., 2006; Gupta et al., 2014). In addition, large-sample datasets allow for model evaluation and analyses across catchments to identify correlations between catchment attributes and model performance (e.g. Donnelly et al. (2016); Konapala et al. (2020); Massmann (2020); David et al. (2022)); thereby not only answering if a model is good but also why (Kirchner, 2006).

However, the relevance of the results of model evaluation and comparison studies is unclear when (streamflow) observation uncertainty is not included in large sample datasets, as is usually the case. As a result the adequacy of hydrological models might be misconstrued. Therefore, a large literature has been devoted on discussing the effect of data quality limitations on hydrological modelling (e.g. Yew Gan et al. (1997); Kirchner (2006); Beven et al. (2011); Kauffeldt et al. (2013); Huang and Bardossy (2020)).

Multiple studies have highlighted the importance of accounting for uncertainties in streamflow observations while conducting hydrological model calibration or evaluation (e.g. McMillan et al. (2010); Coxon et al. (2015); Westerberg et al. (2020)). These studies developed and applied methodologies to determine quantified uncertainty estimates of streamflow observations (overview in McMillan et al. (2012)). Recently Coxon et al. (2020) released the first large-sample dataset that includes quan-

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


tified streamflow observation uncertainty estimates: CAMELS-GB which describes 671 catchments in Great Britain of which 503 gauging stations contain quantified observed streamflow uncertainty information (Coxon et al., 2015).

In this study we investigate the importance of accounting for streamflow observation uncertainty when conducting model evaluation and comparison studies. We created a workflow that assesses the validity of the differences between model simulations in light of observation uncertainty. The generic layout of the workflow allow for assessments that go beyond streamflow in hydrological modelling and is therefore applicable for any field of geoscience where model results are compared against observations with known uncertainty estimates. We extend the study by also considering the effect that the temporal sampling of the simulation and observation time series has on the objective functions used to determine model skill (Clark et al., 2021). The temporal sampling uncertainty is the result of only a few simulation and observation pairs in the streamflow time series having a disproportional effect on the calculated objective function. Clark et al. (2021) identified this as another source of uncertainty that might lead to wrong conclusions based on objective functions that capture streamflow performance as a few data points can heavily influence the results.

The aim of this study is to demonstrate how both observation uncertainty and temporal sampling uncertainty estimates can be included in model evaluation and model comparison studies to provide context to results that is required to draw conclusions based on individual catchments or large-sample catchment datasets.

## 2 Methodology

The graphical workflow in Figure 1 provides an overview of the components of the model experiments and analyses described in the methodology. Figure 1a describes a typical model run with inputs and outputs, Figure 1b describes a classical comparison of objective functions based on (streamflow) observations and simulations, and Figure 1c describes the additional uncertainty analyses introduced in this study.

### 2.1 Model Experiment Inputs

#### 2.1.1 CAMELS-GB dataset

The CAMELS-GB dataset (Coxon et al., 2020; Coxon, 2020) serves as the case study area of the model experiment and contains data (hydro-meteorological timeseries, catchment boundaries and catchment attributes) describing 671 catchments located across Great Britain. The underlying data used to create CAMELS-GB are publicly available and are therefore suitable for evaluating and benchmarking hydrological models as the dataset can be easily extended in the future. A unique feature of the dataset is the availability of quantified streamflow observation uncertainty estimates for the flow percentiles of 503 catchments (see Coxon et al. (2015)). In this study we evaluated 396 of these 503 catchments, as only these contained a complete range of the percentiles of quantified observation uncertainty estimates required for the analyses in Section 2.4.3.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


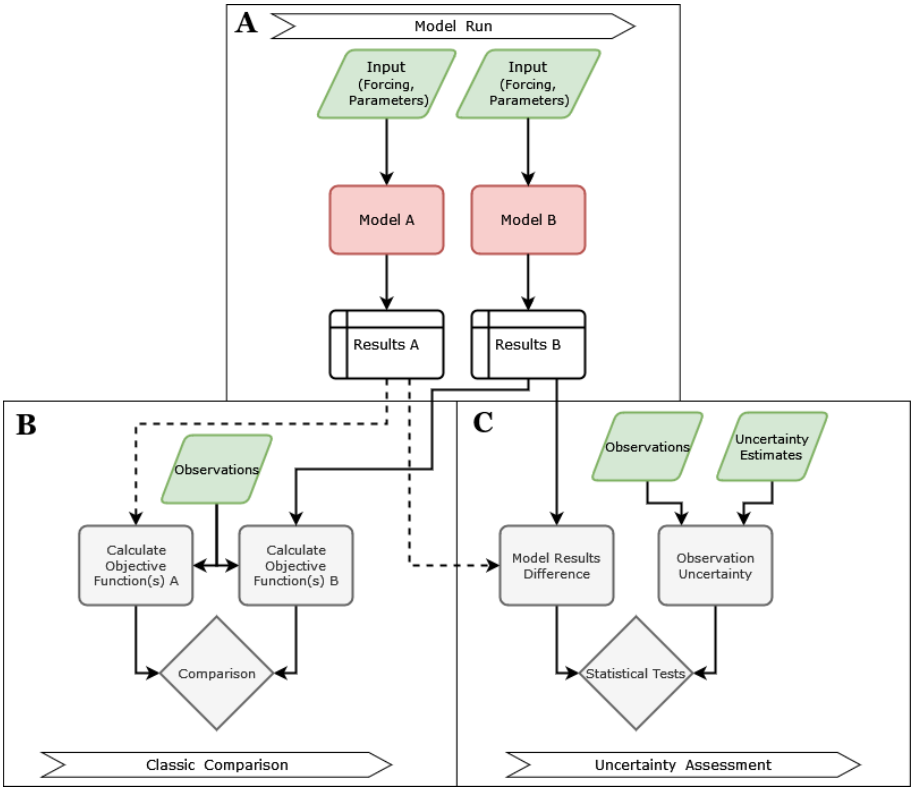


**Figure 1.** Graphical workflow of model experiments and analyses. In green the model experiment inputs, in red the models, in grey the analyses components. Part A describes the model run, part B the classic model comparison that compares objective functions, and part C the workflow for the uncertainty assessment.

### 2.1.2 Meteorological Forcing and Pre-Processing

For consistency we use the same meteorological forcing that was used to create the CAMELS-GB meteorological timeseries and climate indices as input to the hydrological models. This input consists of gridded 1km$^2$ daily meteorological datasets. The meteorological variables used in this study are precipitation (CEH-GEAR; Keller et al. (2015); Tanguy (2021)), reference evaporation (CHESS-PE; Robinson (2020a)), and temperature (CHESSmet; Robinson (2020b)). Scripting used for pre-processing of the data is available in the GitHub repository complementing this study: https://doi.org/10.5281/zenodo.7956488.

### 2.1.3 Streamflow Observations and Quantified Uncertainty Estimates

The streamflow observations in the CAMELS-GB dataset were obtained from the UK National River Flow Archive and are daily values in cubic meters per second. As is common with large-sample datasets several catchments have missing flow data in the time series. These missing values are not taken into account in the analyses of this study.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


A unique aspect of the CAMELS-GB dataset is the inclusion of quantified streamflow observation uncertainty estimates created by Coxon et al. (2015). The uncertainty is quantified by utilizing a large dataset of quality assessed rating curves and stage-discharge measurements. In an iterative process, the mean and variance at each stage point is calculated and subsequently fitted using a LOWESS regression method that defines the rating curve and streamflow uncertainty. By combining the LOWESS curves and variance in a Gaussian Mixture model based on a random draw from the measurement error distribution an estimate of streamflow uncertainty is made, see Coxon et al. (2015) for more information.

## 2.2 Hydrological Models

The selection of the hydrological models in this study is based on the differences in conceptualizations of hydrological processes and calibration routines while being comparable to a certain degree as both are distributed hydrological models that are applicable at fine spatial scale (1km$^2$). In addition, the model selection is in part based on legacy and availability of data (Addor and Melsen (2019)) as well as based on the relevance of the model runs for use in other studies. Below we briefly describe the models. For detailed descriptions the reader is referred to van Verseveld et al. (2022) (wflow_sbm) and Sutanudjaja et al. (2018) (PCR-GLOBWB).

### 2.2.1 wflow_sbm

The wflow_sbm physically based distributed hydrological model (van Verseveld et al., 2022) is based on the Topog_SBM model concept (Vertessy and Elsenbeer, 1999). This concept was developed for small-scale hydrologic simulations. The wflow_sbm model deviates from Topog_SBM by the addition of capillary rise, evapotranspiration and interception losses (Gash model; Gash (1979)), a root water uptake reduction function (Feddes and Zaradny, 1978), glacier and snow processes, and D8 river routing that uses the kinematic wave approximation. The parameter sets (40 in total) are derived from open-source datasets and use pedo-transfer functions to estimate soil properties (see hydroMT software package (Eilander and Boisgontier, 2022). We use the 1 km$^2$ model version that was aggregated from the finest available data scale (90 m).

### 2.2.2 PCR-GLOBWB

The PCR-GLOBWB physically based distributed hydrological model was initially developed for global hydrology and water resources assessments (Sutanudjaja et al., 2018). The PCR-GLOBWB model calculates the water storage in two soil layers, one groundwater layer, and the exchange between the top layer and the atmosphere. The model accounts for water use and return flow determined by water demand. We use the 1 km$^2$ version that is introduced in (Hoch et al., 2023). The model configuration in this study applies the accumulated travel time approximation for river routing.

## 2.3 Model Experiments

We set up two model experiments that cover various purposes for model evaluation. The first experiment, *intra-model*, mimics a model refinement case by evaluating the streamflow estimates of the distributed wflow_sbm hydrological model with and

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


without additional calibration. Hereafter respectively, calibrated and default wflow_sbm. The second experiment, *inter-model*, is a classic model comparison based on the streamflow estimates of the distributed PCR-GLOBWB and wflow_sbm models. These two experiments are in place as we expect that the differences in streamflow simulations between the models of the intra-model experiment are smaller than those of the inter-model experiment and might therefore lead to different conclusions.

#### 2.3.1 PCR-GLOBWB Model Run

Both the wflow_sbm and PCR-GLOBWB hydrological models are setup such as they are typically used in other studies. Therefore, the PCR-GLOBWB model does not require additional calibration using streamflow observations after deriving the parameter set. The model does require an extensive spin-up period to establish semi steady-state conditions at the start of the model run. The model is spun-up 30 years back to back using a single water year climatology that is based on the average values of each calendar-day between 1-10-2000 and 30-09-2007. The water year 2008 is discarded from analyses to avoid overfitting at the start of the evaluation period and the model is evaluated for the water years 2009 - 2015.

#### 2.3.2 Calibrated and Default wflow_sbm Model Runs

The wflow_sbm model is spun-up using the water year 2000 and additionally calibrated using streamflow observations for the water years 2001-2007. Additional calibration is done by optimizing a single parameter using the Kling-Gupta Efficiency Non-Parametric (KGE-NP) objective function (Pool et al. (2018)) based on streamflow observations and simulations at the catchment outlet resulting in a single calibrated parameter set. Imhoff et al. (2020) identified the KsatHorFrac parameter as effective for single parameter value per catchment calibration. KsatHorFrac is an amplification factor of the vertical saturated conductivity that controls the lateral flow in the subsurface. The water year 2008 is discarded from analyses and the model is evaluated for the water years 2009 - 2015. For more information on the effects of calibration, the reader is referred to Aerts et al. (2022), Section 3.1 and Figure 3. The default wflow_sbm model run sets the KsatHorFrac parameter value to the default value of 100.

#### 2.3.3 eWaterCycle

This study is conducted using the eWaterCycle platform (Hut et al., 2022). eWaterCycle is a community driven platform for the running of hydrological model experiments. All components that are required to run hydrological models are FAIR by design (Wilkinson et al., 2018). This is achieved by versioning models and datasets and creating workflows that are reproducible. Therefore, the platform is suitable for conducting benchmark experiments. The model simulations were conducted on the Dutch supercomputer Snellius to ensure faster model run time. We created example notebooks that use the eWatercycle platform on cloud computing infrastructure: https://doi.org/10.5281/zenodo.7956488.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


## 2.4 Analyses

### 2.4.1 Model Evaluation

The hydrological model results are evaluated using the Kling-Gupta efficiency non-parametric (KGE-NP, Pool et al. (2018)) objective function. This efficiency metric deviates from the more commonly used Kling-Gupta efficiency (KGE, Gupta et al. (2009)) by calculating the Spearman rank correlation and the normalized-flow-duration curve instead of the Pearson correlation and variability bias. Values range from -$\infty$ to 1 (perfect score). In addition to the KGE-NP metric, we consider the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe (1970)) to demonstrate the sensitivity of the results towards the selection of objective function. We include the KGE-NP, KGE, modified KGE (Kling et al. (2012)),and the NSE objective functions in the data repository for completeness and future reference.

### 2.4.2 Temporal Sampling Uncertainty

Time series of observations and simulations are not infinitely long and might contain outliers. Therefore, there is uncertainty surrounding the sampling of time series on which model skill is based. The temporal sampling uncertainty is the result of a few observation and simulation pairs in the (streamflow) time series having a disproportionate effect on the calculated objective function that is used to determine model performance. These observation and simulation pairs are viewed as outliers (e.g. extreme high flows) and when present these should be evaluated before drawing conclusions on model performance.

To quantify the temporal sampling uncertainty of the KGE-NP and NSE objective functions we applied the methodology of Clark et al. (2021). This method combines bootstrapping of shorter time intervals of the observation and simulation time series (Efron, 1979) and the jacknife-after-bootstrap (Efron and Tibshirani, 1986) method, the shuffling of these intervals, to calculate the standard error and the tolerance interval of the temporal sampling uncertainty. This follows the methodology of (Clark et al., 2021). We extended the GUMBOOT package (Clark et al., 2021) by adding the KGE-NP metric. The analysis of the temporal sampling uncertainty of each objective function is based on the tolerance interval (95th - 5th percentiles) of the jackknife and bootstrap methods. The tolerance intervals of the models corresponding to each model experiment are averaged for each catchment and is referred to as temporal sampling uncertainty.

### 2.4.3 Streamflow Observation Uncertainty

The results are analyzed based on three flow categories, similar to Coxon et al. (2015), namely low flow, average flow, and high flow conditions. Low flow is based on the values between the 5th and 25th percentiles, average flow on the 25th and 75th percentiles, and high flow on the 75th and 95th percentiles of observed streamflow at the catchment outlet. Not all percentiles are included for the low and high flow categories due to limited data availability on quantified streamflow observation uncertainty. Following the creation of flow categories based on the percentiles of observed flow, the simulated streamflow results are divided into flow categories by matching the time steps (dates).

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


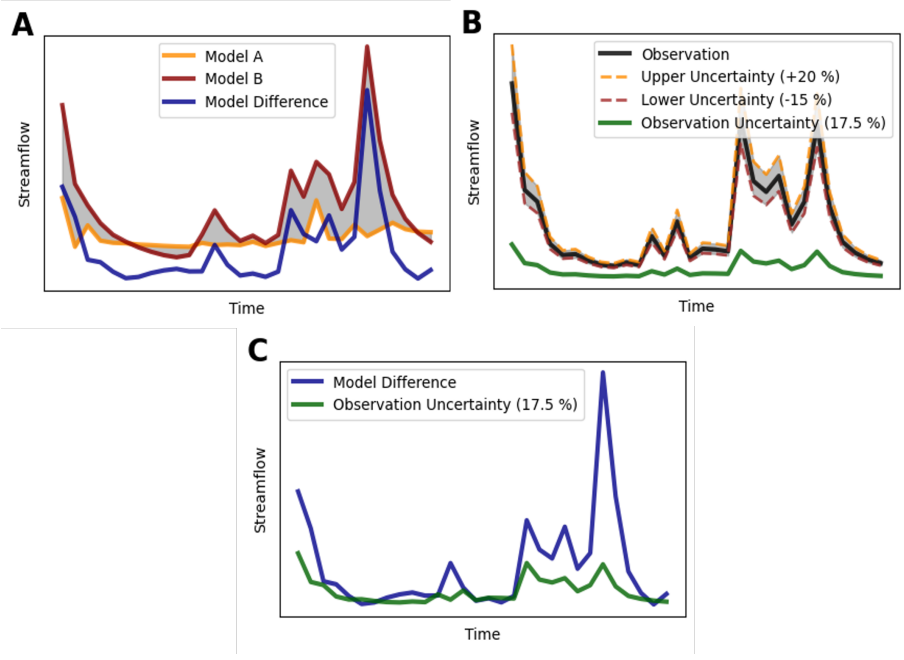


**Figure 2.** Example hydrographs of the streamflow observation uncertainty analysis method. (A) Calculation of the absolute difference (blue) between model simulations (red and orange). (B) Calculation of streamflow observation uncertainty in m3/s (green). Dashed lines indicating upper and lower bounds expressed as percentages of observation uncertainty that are averaged and multiplied with the observations (black). (V) Resulting time series, with, in blue, the absolute difference between model simulations and, in green, the averaged observation uncertainty in m3/s.

Next, the method illustrated in Figure 1C is applied to each flow category, catchment, and objective function. First the absolute difference of the model simulations is calculated as shown in the example hydrograph in Figure 2A. The quantified streamflow observation uncertainty estimates of the CAMELS-GB dataset contain upper and lower bounds per percentile (5,25,75,95). First the percentile boundaries of each flow category are averaged (e.g. 5th and 25th percentiles for the low flow category) resulting in the orange and red dashed lines in Figure 2B. Next, the average of upper and lower bounds are taken and multiplied by the observations to create the quantified streamflow observation estimate time series in m3/s (green line). These bounds differ in values (e.g, +20% and -15%) as the uncertainty distributions are not symmetrical.

A t-test is performed using the time series in Figure 2C with a 0.05 significance level to determine if the observation uncertainty is greater than the model simulation difference. When this is the case it is not possible to be conclusive on which model is best performing.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


## 3 Results

Firstly, we present an overview of streamflow simulation based model performance captured by the KGE-NP and NSE objective functions at the catchment outlets for each model experiment. Secondly, we show the distributions of the difference between models per experiment in objective function and the temporal sampling uncertainty for both objective functions. Thirdly, we show for each flow category the percentage of days that the observation uncertainty is greater than the differences between model simulations of each experiment. A t-test determines if the model simulation difference time series are significantly greater than the observation uncertainty time series.

The calibration results of the wflow_sbm model are available in Appendix A1. Appendix A2 contains an overview of the distributions of the GUMBOOT (sampling uncertainty) results for each objective function.

### 3.1 Streamflow Based Model Performance

The model performance results are based on the streamflow estimates and streamflow observations at 396 catchment outlets and are shown as cumulative distributions functions in Figure 3. In general the results show, to a certain degree, comparable skill in capturing observed streamflow by both hydrological models. Note, that the different objective functions in Figures 3ab are not directly comparable ((Knoben et al., 2019)).

The inter-model comparison experiment results of the calibrated wflow_sbm & PCR-GLOBWB models show a large difference between the KGE-NP distributions of the models above a KGE-NP of 0.25 (Figure 3a). The KGE-NP median of the calibrated wflow_sbm is 0.77 compared to a median value of 0.43 for PCR-GLOBWB. Larger differences between distributions and, in general, lower values are found based on the NSE metric in Figure 3b. The large differences are in part due to the additional calibration of the wflow_sbm model. Another contributing factor is expected to be the difference in river routing, kinematic wave used by wflow_sbm and simple accumulation travel time by PCR-GLOBWB. The differences between objective functions can be explained by the KGE-NP function focusing more on the baseflow component while the NSE objective function focuses more on average and peak flow.

The intra-model evaluation experiment results capture the effect of additional calibration of the wflow_sbm model. This is shown by the default and calibrated wflow_sbm model distributions with median values of 0.65 and 0.77 respectively. The added value of calibration is less pronounced for the NSE results in Figure 3b as the model calibration routine only optimizes for the KGE-NP objective function. Here, the median values are lower at 0.25 for the default and 0.50 for the calibrated wflow_sbm models. The differences between distributions of each objective function establishes the importance of reporting multiple performance metrics. Overall, the differences are larger for the inter-model comparison experiment, smaller for the intra-model evaluation experiment, and more pronounced for the NSE than the KGE-NP objective function.

### 3.2 Temporal Sampling Uncertainty

The distributions of the difference between objective functions of the models for both experiments and the temporal sampling uncertainty are shown in Figure 4. The objective function difference distribution shows a larger spread in values for the inter-

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


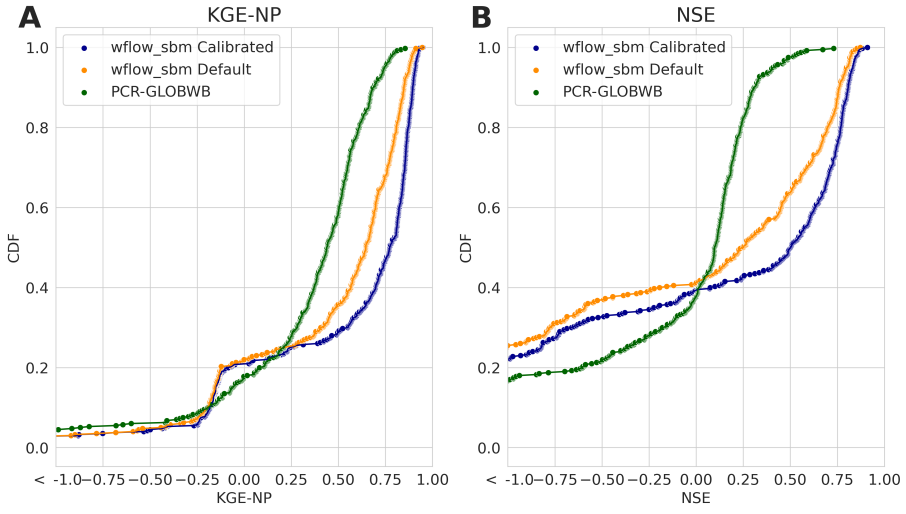


**Figure 3.** Cumulative distribution function (CDF) plots of objective functions based on streamflow estimate and observation at the catchment outlet. With in blue the additionally calibrated wflow_sbm model, in orange the default wflow_sbm model, and in green the PCR-GLOBWB model. (A) The Kling-Gupta Efficiency non-parametric (KGE-NP) objective function. (B) The Nash Sutcliffe Efficiency (NSE) objective function. The results show closer agreement for objective function values of the intra-model evaluation than the inter-model comparison

model comparison (KGE-NP median of 0.48) than for the intra-model evaluation (KGE-NP median of 0.07). This is to lesser extent the case for the temporal sampling uncertainty distributions. Both model experiments show higher values for the objective function difference and temporal sampling uncertainty of the NSE than the KGE-NP objective function. This demonstrates a strong sensitivity of results towards objective function selection. For both objective functions very large values of differences are present in the negative domain. The relevance of which is debatable as for example (Knoben et al., 2019) pointed out that a KGE value of -0.42 and NSE of 0 corresponds to taking the mean of the observations.

Next, the catchment simulations that contain greater sampling uncertainty than the difference in objective functions are identified (Table 1). Of the 398 catchment simulations under consideration this is the case for 53 catchments based on the KGE-NP objective function and 86 catchments based on the NSE objective function of the inter-model comparison. The intra-model evaluation contains more cases as the objective function differences are lower while the sampling uncertainty is similar. This results in 210 catchment simulations that contain greater sampling uncertainty than objective function differences for the KGE-NP objective function and 288 catchments for the NSE objective function. These results demonstrate that in many catchments data points in the tails of the probability distribution of the squared errors between model simulations and observations (outliers) heavily influence the objective function. It is difficult to determine whether the objective function differences are the result of modelling differences or mainly due to temporal sampling. Therefore further research is required that determines the validity of these data points that heavily influence the objective function before drawing conclusions on model performance.

The spatial distribution of the temporal sampling uncertainty results in Figure 5 show clusters of high sampling uncertainty for all model experiments and objective functions in the South of Great-Britain. This is most likely due to the presence of chalk

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


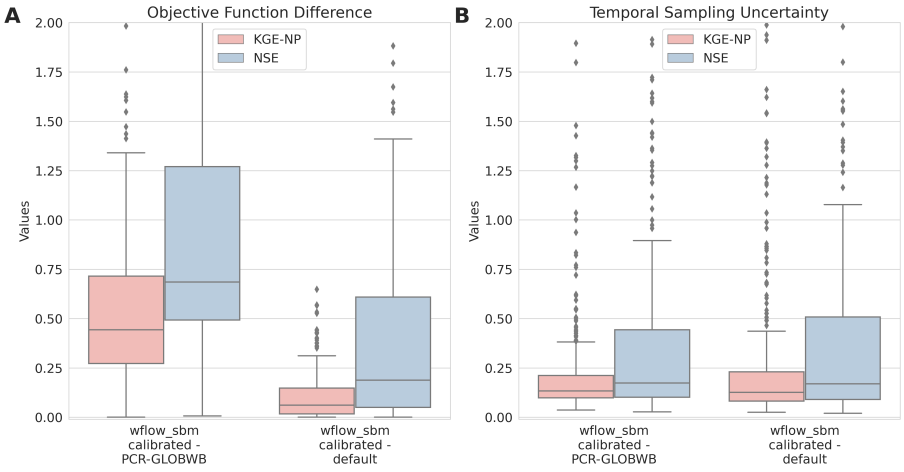


**Figure 4.** (A) Box-plot of the objective function difference between the inter-model comparison models (wflow_sbm calibrated and PCR-GLOBWB) and the intra-model evaluation models (wflow_sbm calibrated and default). (B) Box-plot of the temporal sampling uncertainty (average tolerance interval) of both model experiments. The KGE-NP objective function is shown in red and the NSE objective function in blue.

**Table 1.** Number of catchment simulations per model experiment and objective function for which temporal sampling uncertainty (average tolerance interval) is larger than the difference in objective function. 398 catchment simulations are considered.

| Model Experiment | Models | Objective Function | Sampling Uncertainty > Objective Function Difference |
|---|---|---|---|
| Inter-model Comparison | wflow_sbm & PCR-GLOBWB | KGE-NP | 53 |
| Inter-model Comparison | wflow_sbm & PCR-GLOBWB | NSE | 86 |
| intra-model Evaluation | wflow_sbm calibrated & default | KGE-NP | 210 |
| intra-model Evaluation | wflow_sbm calibrated & default | NSE | 288 |

geology that is known to cause difficulties for estimating streamflow using hydrological models. The inter-model comparison results in Figures 5ab show that there is agreement on where high sampling uncertainty (>0.4) occurs. This is more so the case for the NSE- than the KGE-NP objective function. The intra-model evaluation experiment results in Figures 5cd show more agreement on occurrences than the inter-model comparison. These results show only clusters of objective function differences greater than sampling uncertainty in the West and North of Great-Britain. In addition, more catchments contain very high sampling uncertainty (>0.4) indicating that the averaging of the tolerance interval reduces the sampling uncertainty more for the inter-model comparison.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


**Table 2.** Result of t-test with an statistical significance level of 0.05 determining if the observation uncertainty time series is larger than the simulation time series for each flow category and model experiment. 398 catchments are considered.

| Model Experiment | Models | Flow Category | Observation Uncertainty >Simulation Difference (p-value 0.05) |
|---|---|---|---|
| Inter-model Comparison | wflow_sbm & PCR-GLOBWB | Low Flow | 6 |
| Inter-model Comparison | wflow_sbm & PCR-GLOBWB | Average Flow | 4 |
| Inter-model Comparison | wflow_sbm & PCR-GLOBWB | High Flow | 3 |
| intra-model Evaluation | wflow_sbm calibrated & default | Low Flow | 116 |
| intra-model Evaluation | wflow_sbm calibrated & default | Average Flow | 114 |
| intra-model Evaluation | wflow_sbm calibrated & default | High Flow | 138 |

### 3.3 Streamflow Observation Uncertainty

The observation uncertainty percentages per flow category and the percentage of days that the observation uncertainty is greater than the model simulation differences (see Figure 2c) are shown in Figure 6. The observation uncertainty percentages in Figure 6A indicate high percentages of uncertainty throughout the case study area with median values of 19.85 (low flow), 15.52 (average flow), and 12.18 (high flow). All flow categories contain outliers of more than 50 % observation uncertainty. Of interest is that the uncertainty percentages are highest for the low flow category while the agreement between model simulations is highest for this flow category. This is shown by the lower percentages of days that the observation uncertainty is greater than the simulation differences between models in Figure 6b. In addition, smaller simulation differences (intra-model comparison) result in more percentages of days of observation uncertainty values surpassing simulation difference values.

Next, we applied a t-test to determine in which catchments the observation uncertainty is statistically larger than the differences between simulations for each flow category and experiment (Table 2). For the inter-model comparison experiment we find that this is the case for 6 catchments of the low flow, 4 catchments of the average flow, and 3 catchments of the high flow category. These are low values but can potentially influence large-sample scale conclusions. After exclusion of these catchments we can conclude that the model comparison is not heavily influenced by the observation uncertainty. The smaller differences between simulation in the intra-model experiment result in many catchment simulations for which additional calibration does not significantly lead to improvements of streamflow estimates in light of observation uncertainty. This is the case for 116 of the low flow, 114 of the average flow, and 138 catchments of the high flow category. The differences in simulations are more substantial for the inter-model comparison as only a few catchments per flow category are smaller than the observation uncertainty.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


## 4 Discussion

This study highlights the importance of taking into account streamflow observation uncertainty and temporal sampling uncertainty when evaluating or comparing hydrological models based on time series. We acknowledge that these are not the only sources of uncertainty as there is uncertainty in model inputs, model structure, parameter sets, initial or boundary conditions, and more (e.g. Renard et al. (2010); Dobler et al. (2012); Hattermann et al. (2018); Moges et al. (2021)). A full uncertainty analysis of the complete modelling chain is needed for a complete picture (Beven and Freer (2001); Pappenberger and Beven (2006); Beven (2006)). The uncertainty sources assessed in this paper are only one part, but an often overlooked one.

### 4.1 From temporal sampling uncertainty to certainty

The temporal sampling uncertainty assessed in Section 3.2 is governed by outliers in the probability distribution of the squared errors between model simulations and observations. High values of temporal sampling uncertainty indicate that certain data points have an exceptionally large effect on the objective function. It is therefore important to investigate the validity of these data points as measurement error or model error might misconstrue the actual model performance. For example, the spatial distribution of the results showed agreement on high sampling uncertainty clustered in the South of Great-Britain. This region contains the karst (chalk) geology that is known to be difficult to model correctly (Hartmann et al. (2014)). Further inspection of the streamflow observations at the catchment outlets did not show unexpected outliers that might indicate measurement error. It is therefore likely that differences between observations and simulations are large and inconsistent and might not only be influenced by how the time series are sampled. Through the detailed inspection of the time series we can deem with a higher degree of certainty that the results are not unjustly influenced by temporal sampling.

In addition, we compared the distributions of the sampling uncertainty results of each model run and objective function in Appendix A2 to those presented for the VIC model using the large-sample CAMELS-US dataset by Clark et al. (2021). The distributions of this study are similar for each model experiment and objective function and of similar magnitude to those of Clark et al. (2021). Therefore, the same conclusion is valid for both studies in that care should be taken before drawing conclusions at the large-sample scale. This is especially the case for the identified catchments in this study that contain sampling uncertainty values greater than the difference in objective functions between two model simulations.

### 4.2 Why we should account for observation uncertainty

The results of this study demonstrate that (streamflow) observation uncertainty is important to consider when comparing or evaluating hydrological models. If the difference between model simulations is within the uncertainty bounds of the observation uncertainty it is not possible to draw conclusions on best performing model simulations. The intra-model experiment shows that smaller differences between models such as changes made to model structure, inputs, or parameterization and calibration result in more of these occurrences. This is the case for 123 catchments based on the average of all flow categories. This does not mean that the incremental improvements to the model structure are not important, but it does show that they might not be as relevant as expected in light of observation uncertainty. The inter-model comparison contained only 3 to 6 catchments

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


(depending on the flow category) of significantly higher observation uncertainty than simulation differences. We recommend that these catchment simulations are removed from benchmarks or model comparisons when this is the case.

In this study we used the limits of streamflow observation uncertainty at the catchment outlets as described in the CAMELS-GB dataset. Besides the limitations of the quantification of the observation uncertainty itself, this study is limited by the availability of only the uncertainty bounds of uncertainty. If we had ideal data available we would use the standard deviations of the observation uncertainty distributions as these are more conservative estimates. A repeat of the observation uncertainty analysis using these estimates would result in less catchment simulations showing higher observation uncertainty than the differences between model simulations.

### 4.3 Moving towards standardized benchmark procedures

We introduced a method that accounts for streamflow observation uncertainty which is kept as generic and easy to implement as possible. The generality ensures broader applicability in hydrology and geosciences. The method is applicable for any state or flux for which observation time series including uncertainty estimates are available. In the absence of uncertainty estimates, one might use this method in combination with multiple evaluation products. A rough estimate of uncertainty can be based on the probability density distribution of multiple observation time series. The ease of implementation is key as it more likely to be adopted by other studies and to be part of standardized benchmarks.

Benchmarks are valuable for model evaluation to support interpretation of model performance in other studies (Seibert, 2001; Schaefli and Gupta, 2007; Pappenberger et al., 2015; Seibert et al., 2018). The inclusion of observation and temporal sampling uncertainty in the benchmark procedure not only provides a better indication of the relevance of the differences between benchmark results, but also helps detecting benchmark samples that should be treated with care or may need to be excluded from the benchmark. We therefore advocate the reporting of both types of uncertainty in benchmark procedures. Reporting can be further improved by separating flow conditions in the case of streamflow as observation uncertainties differ. The additional information through flow separation can be used to support hypotheses related to connections between streamflow simulations and hydrological process descriptions. This distils into reporting more meta-data with model outputs in a standardized manner.

For statistical and model benchmarks to be standardized it is necessary that the community agrees on best practices and provides a template for benchmark experiments and reporting and storage (Hoch and Trigg (2019)). Standardized benchmark procedures will increase the longevity of model benchmark results for future research. Standardization will also reduce redundant work as less model runs are required. This has the benefit of stimulating more time spent on novel research than data intensive studies (Jain et al., 2022). Standardized benchmark templates should encompass multiple objective function, as is reconfirmed by the sensitivity of results to objective function selection in this study, and workflows for the evaluation of multiple states and fluxes.

Here, we make an effort to standardize the workflow by firstly using the same meteorological forcing data and streamflow observations that were used to create the CAMELS-GB dataset for consistency and secondly through the creation of reproducible workflows using eWaterCycle (Hut et al., 2022). We use eWaterCycle to show how benchmark studies can be done in a reproducible manner using high level readable code. Platforms like eWaterCycle should host standardized benchmark

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


procedures to achieve the benefits outlined above. With this study we aim to set first steps by providing documented example notebooks of the scripting (https://doi.org/10.5281/zenodo.7956488). This can be viewed as a template for a benchmark procedure when studying the difference in hydrological model performance in the light of observation uncertainty and temporal sampling uncertainty. To facilitate comparisons between different studies we encourage the hydrological community when doing benchmark studies to either use, or add to the collection of, community standard benchmark templates. Future work should extent the benchmark procedure to include evaluation of multiple states and fluxes.

## 5 Conclusions

We set out this study to highlight the importance of including streamflow observation uncertainty and temporal sampling uncertainty when conducting hydrological model evaluations or model comparisons based on large-sample hydrology dataset. By developing a generic and easy to implement method we demonstrated how these uncertainties can be included in benchmark procedures. The scripting accompanying this study is easily adaptable to other case study areas, hydrological models, and forcing inputs due to the implementation in eWaterCycle.

We demonstrated the methodology through two experiments. The first epxeriment is an inter-model comparison comparison experiment of the wflow_sbm and PCR-GLOBWB hydrological models. The second experiment mimics a model refinement approach by intra-model evaluation that assesses the benefits of additional calibration based on streamflow observations of the wflow_sbm model.

The main findings of these experiments are that for the sampling uncertainty assessment the intra-model evaluation experiment simulations of 210 (KGE-NP) and 288 (NSE) out of 398 catchments contain higher sampling uncertainty than the difference in objective functions. For the inter-model comparison experiment simulations these are 53 (KGE-NP) and 86 (NSE) catchments out of 398. Based on these catchments it is difficult to draw conclusions as to which model is best performing based on streamflow at the catchment outlet before further investigating the validity of the data points causing the sampling uncertainty. The high number of occurrences establish and highlight the importance of reporting sampling uncertainty.

For the observation uncertainty assessment, the intra-model evaluation experiment shows, depending on the flow category, between 114 and 138 catchment simulations with statistically higher streamflow observation uncertainty than differences between model simulations. Hence, no conclusions can be drawn on the better performing model based on these catchments. Lower values of between 3 and 6 catchment simulations are found for the inter-model comparison experiment. These should be reported and excluded from benchmarking. Given that the number of catchments is low, large-sample scale conclusions are not as strongly affected by the streamflow observation uncertainty.

These experiments demonstrated the importance of not accepting the output of benchmark efforts on face value when uncertainties between models and model observations are not acctounted for explicitly. Implementing the proposed method in standardized benchmark procedures will lead to more robust benchmarking results.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


*Code availability.* https://github.com/jeromaerts/CAMELS-GB_Comparison_Uncertainty, https://doi.org/10.5281/zenodo.7956488

*Author contributions.* JPMA wrote the publication. JPMA, JMH, and RWH, conceptualized the study. JPMA, JMH, RWH, NCvdG, GC developed the methodology. JPMA, JMH, conducted the analyses. JMH, RWH, NCvdG, GC did internal reviews. RWH, NCvdG are PIs of the eWaterCycle project.

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

*Acknowledgements.* This work has received funding from the Netherlands eScience Center (NLeSC) under file number 027.017.F0. We would like to thank the research software engineers (RSEs) at NLeSC who co-built the eWaterCycle platform and Surf for providing computing infrastructure. Gemma Coxon was supported by a UKRI Future Leaders Fellowship award [MR/V022857/1].

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


## References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resources Research, 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 65, 712–725, https://doi.org/10.1080/02626667.2019.1683182, 2020.

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., and Hazenberg, P.: Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model, Hydrology and Earth System Sciences, 26, 4407–4430, https://doi.org/10.5194/hess-26-4407-2022, 2022.

Andréassian, V., Hall, A., Chahinian, N., and Schaake, J.: Introduction and synthesis: Why should hydrologists work on a large number of basin data sets?, in: Large sample basin experiments for hydrological parametrization : results of the models parameter experiment-MOPEX. IAHS Red Books Series n° 307, pp. 1–5, AISH, https://hal.inrae.fr/hal-02588687, 2006.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Beven, K., Smith, P. J., and Wood, A.: On the colour and spin of epistemic error (and what we might do about it), Hydrology and Earth System Sciences, 15, 3123–3133, https://doi.org/10.5194/hess-15-3123-2011, 2011.

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/10.1029/2020WR029001, 2021.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531–5546, https://doi.org/10.1002/2014WR016532, 2015.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Coxon, G.;Addor, N. J. J. M. J. N. R. M. E. T. R.: Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB), https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020.

David, P. C., Chaffe, P. L. B., Chagas, V. B. P., Dal Molin, M., Oliveira, D. Y., Klein, A. H. F., and Fenicia, F.: Correspondence Between Model Structures and Hydrological Signatures: A Large-Sample Case Study Using 508 Brazilian Catchments, Water Resources Research, 58, e2021WR030 619, https://doi.org/10.1029/2021WR030619, 2022.

Dobler, C., Hagemann, S., Wilby, R. L., and Stötter, J.: Quantifying different sources of uncertainty in hydrological projections in an Alpine watershed, Hydrology and Earth System Sciences, 16, 4343–4360, https://doi.org/10.5194/hess-16-4343-2012, 2012.

Donnelly, C., Andersson, J. C., and Arheimer, B.: Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe, Hydrological Sciences Journal, 61, 255–273, https://doi.org/10.1080/02626667.2015.1027710, 2016.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


Efron, B.: Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7, 1–26, https://doi.org/10.1214/aos/1176344552, 1979.

Efron, B. and Tibshirani, R.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science, 1, 54–75, https://doi.org/10.1214/ss/1177013815, 1986.

Eilander, D. and Boisgontier, H.: hydroMT, https://doi.org/10.5281/zenodo.6107669, 2022.

Feddes, R. A. and Zaradny, H.: Model for simulating soil-water content considering evapotranspiration — Comments, Journal of Hydrology, 37, 393–397, https://doi.org/10.1016/0022-1694(78)90030-6, 1978.

Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment, preprint, Catchment hydrology/Modelling approaches, https://doi.org/10.5194/hess-2022-245, 2022.

Gao, H., Sabo, J. L., Chen, X., Liu, Z., Yang, Z., Ren, Z., and Liu, M.: Landscape heterogeneity and hydrological processes: a review of landscape-based hydrological models, Landscape Ecology, 33, 1461–1480, https://doi.org/10.1007/s10980-018-0690-4, 2018.

Gash, J. H. C.: An analytical model of rainfall interception by forests, Quarterly Journal of the Royal Meteorological Society, 105, 43–55, https://doi.org/10.1002/qj.49710544304, 1979.

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, 2009.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., and Weiler, M.: Karst water resources in a changing world: Review of hydrological modeling approaches, Reviews of Geophysics, 52, 218–242, https://doi.org/10.1002/2013RG000443, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2013RG000443, 2014.

Hattermann, F. F., Vetter, T., Breuer, L., Su, B., Daggupati, P., Donnelly, C., Fekete, B., Flörke, F., Gosling, S. N., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Müller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T., and Krysnaova, V.: Sources of uncertainty in hydrological climate impact assessment: a cross-scale study, Environmental Research Letters, 13, 015 006, https://doi.org/10.1088/1748-9326/aa9938, publisher: IOP Publishing, 2018.

Hoch, J. M. and Trigg, M. A.: Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models, Environmental Research Letters, 14, 034 001, https://doi.org/10.1088/1748-9326/aaf3d3, publisher: IOP Publishing, 2019.

Hoch, J. M., Sutanudjaja, E. H., Wanders, N., van Beek, R. L. P. H., and Bierkens, M. F. P.: Hyper-resolution PCR-GLOBWB: opportunities and challenges from refining model spatial resolution to 1 km over the European continent, Hydrology and Earth System Sciences, 27, 1383–1401, https://doi.org/10.5194/hess-27-1383-2023, publisher: Copernicus GmbH, 2023.

Huang, Y. and Bardossy, A.: Impacts of Data Quantity and Quality on Model Calibration: Implications for Model Parameterization in Data-Scarce Catchments, Water, 12, 2352, https://doi.org/10.3390/w12092352, 2020.

Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., Aerts, J., Albers, T., Alidoost, F., Andela, B., Camphuijsen, J., Dzigan, Y., van Haren, R., Hutton, E., Kalverla, P., van Meersbergen, M., van den Oord, G., Pelupessy, I., Smeets, S., Verhoeven, S., de Vos, M., and Weel, B.: The eWaterCycle platform for open and FAIR hydrological collaboration, Geoscientific Model Development, 15, 5371–5390, https://doi.org/10.5194/gmd-15-5371-2022, 2022.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., and Weerts, A. H.: Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution Distributed Hydrologic Modeling: An Example for the Rhine River, Water Resources Research, 56, e2019WR026 807, https://doi.org/10.1029/2019WR026807, 2020.

Jain, S., Mindlin, J., Koren, G., Gulizia, C., Steadman, C., Langendijk, G. S., Osman, M., Abid, M. A., Rao, Y., and Rabanal, V.: Are We at Risk of Losing the Current Generation of Climate Researchers to Data Science?, AGU Advances, 3, e2022AV000 676, https://doi.org/10.1029/2022AV000676, 2022.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845–2857, https://doi.org/10.5194/hess-17-2845-2013, 2013.

Keller, V. D. J., Tanguy, M., Prosdocimi, I., Terry, J. A., Hitt, O., Cole, S. J., Fry, M., Morris, D. G., and Dixon, H.: CEH-GEAR: 1 km resolution daily and monthly areal rainfall estimates for the UK for hydrological and other applications, Earth System Science Data, 7, 143–155, https://doi.org/10.5194/essd-7-143-2015, 2015.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, https://doi.org/10.1029/2005WR004362, 2006.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

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, publisher: Copernicus GmbH, 2019.

Konapala, G., Kao, S.-C., Painter, S. L., and Lu, D.: Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US, Environmental Research Letters, 15, 104 022, https://doi.org/10.1088/1748-9326/aba927, 2020.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, https://eartharxiv.org/repository/view/3345/, 2022.

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A., and Reaney, S. M.: Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain, Hydrology and Earth System Sciences, 23, 4011–4032, https://doi.org/10.5194/hess-23-4011-2019, 2019.

Massmann, C.: Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments, Hydrological Processes, 34, 4–20, https://doi.org/10.1002/hyp.13566, 2020.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrological Processes, 24, 1270–1284, https://doi.org/10.1002/hyp.7587, 2010.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrological Processes, 26, 4078–4111, https://doi.org/10.1002/hyp.9384, 2012.

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resources Research, 53, 8020–8040, https://doi.org/10.1002/2017WR020401, 2017.

Moges, E., Demissie, Y., Larsen, L., and Yassin, F.: Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis, Water, 13, 28, https://doi.org/10.3390/w13010028, 2021.

Nash, J. E. and Sutcliffe, J. V.: 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.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resources Research, 42, https://doi.org/10.1029/2005WR004820, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004820, 2006.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, Journal of Hydrology, 522, 697–713, https://doi.org/10.1016/j.jhydrol.2015.01.024, 2015.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrological Sciences Journal, 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., and Samaniego, L.: Diagnostic Evaluation of Large-Domain Hydrologic Models Calibrated Across the Contiguous United States, Journal of Geophysical Research: Atmospheres, 124, 13 991–14 007, https://doi.org/10.1029/2019JD030767, 2019.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, https://doi.org/10.1029/2009WR008328, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2009WR008328, 2010.

Robinson, E.L.;Blyth, E. D.-P. E. A.: Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2017) [CHESS-met], https://doi.org/10.5285/2ab15bf0-ad08-415c-ba64-831168be7293, 2020a.

Robinson, E.L.;Blyth, E. D.-P. E. A.: Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2017) [CHESS-PE], https://doi.org/10.5285/9116e565-2c0a-455b-9c68-558fdd9179ad, 2020b.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, 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/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/10.1002/hyp.11476, 2018.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geoscientific Model Development, 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, 2018.

Tanguy, M.;Dixon, H. I. D. V.: Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019) [CEH-GEAR], https://doi.org/10.5285/dbf13dd5-90cd-457a-a986-f2f9dd97e93c, 2021.

van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Bouaziz, L., Eilander, D., Hegnauer, M., ten Velden, C., and Russell, B.: Wflow_sbm v0.6.1, a spatially distributed hydrologic model: from global data to local applications, preprint, Hydrology, https://doi.org/10.5194/gmd-2022-182, 2022.

Vertessy, R. A. and Elsenbeer, H.: Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization, Water Resources Research, 35, 2173–2187, https://doi.org/10.1029/1999WR900051, 1999.

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., and Seibert, J.: Hydrological model calibration with uncertain discharge data, Hydrological Sciences Journal, 0, 1–16, https://doi.org/10.1080/02626667.2020.1735638, 2020.

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., and Dumontier, M.: A design framework and exemplar metrics for FAIRness, Scientific Data, 5, 180 118, https://doi.org/10.1038/sdata.2018.118, 2018.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


520 Yew Gan, T., Dlamini, E. M., and Biftu, G. F.: Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling, Journal of Hydrology, 192, 81–103, https://doi.org/10.1016/S0022-1694(96)03114-9, 1997.

## 1 wflow_sbm calibration

### A1 Temporal sampling uncertainty distributions of the objective functions

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


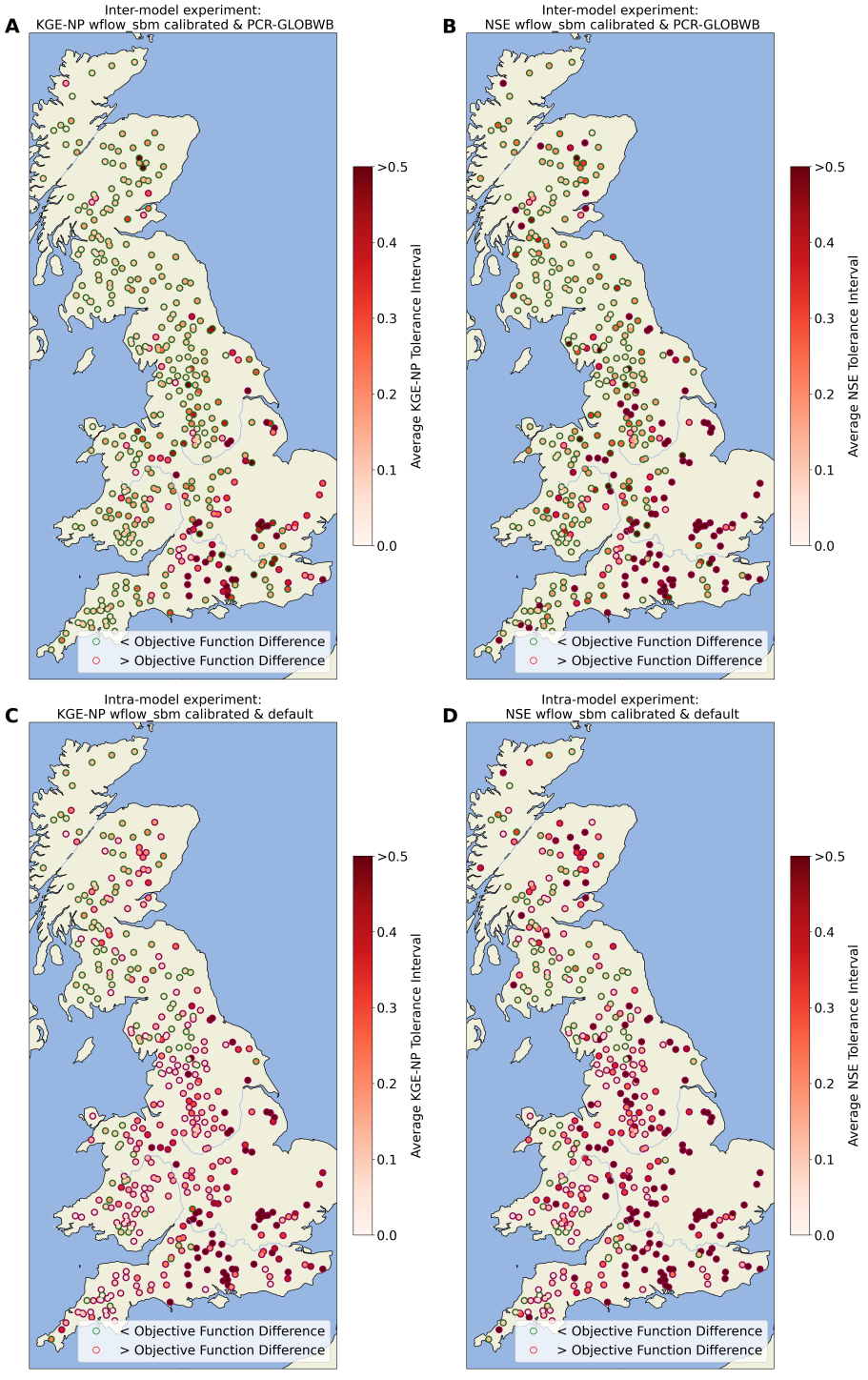


**Figure 5.** Spatial distribution of the temporal sampling uncertainty analyses results showing the average tolerance interval of temporal sampling uncertainty per objective function from white to dark red. Red circles indicate temporal sampling uncertainty larger than objective function difference and green circles indicate sampling uncertainty smaller than objective function difference. (A) Inter-model comparison experiment (wflow_sbm and PCR-GLOBWB) KGE-NP objective function. (B) Inter-model comparison experiment NSE objective function. (C) intra-model evaluation experiment (wflow_sbm calibrated and default) KGE-NP objective function. (D) intra-model evaluation experiment NSE objective function.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


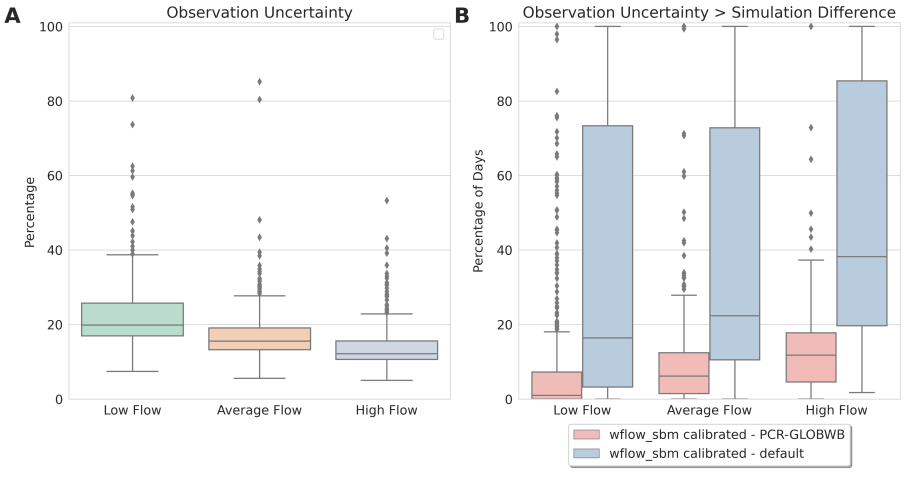


**Figure 6.** (A) Distributions of observation uncertainty percentages per flow category of 398 catchments. (B) The percentage of days that the streamflow observation uncertainty is larger than the difference in streamflow simulation per flow category of 398 catchments. With in red the inter-model experiment (calibrated wflow_sbm and PCR-GLOBWB) and in blue the intra-model evaluation experiment (calibrated and default wflow_sbm).

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


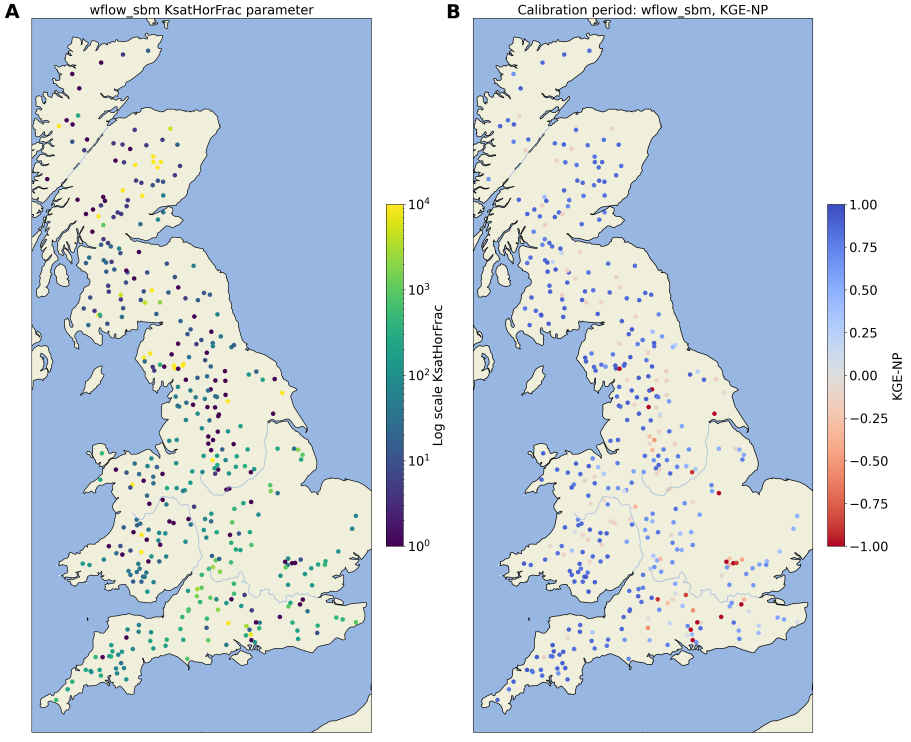


**Figure A1.** (A) Spatial distribution of the best performing KsatHorFrac calibration parameter of the wflow_sbm model based on additional calibration on streamflow observations. (B) Spatial distribution of the KGE-NP objective function based on the calibration period of the wflow_sbm model.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


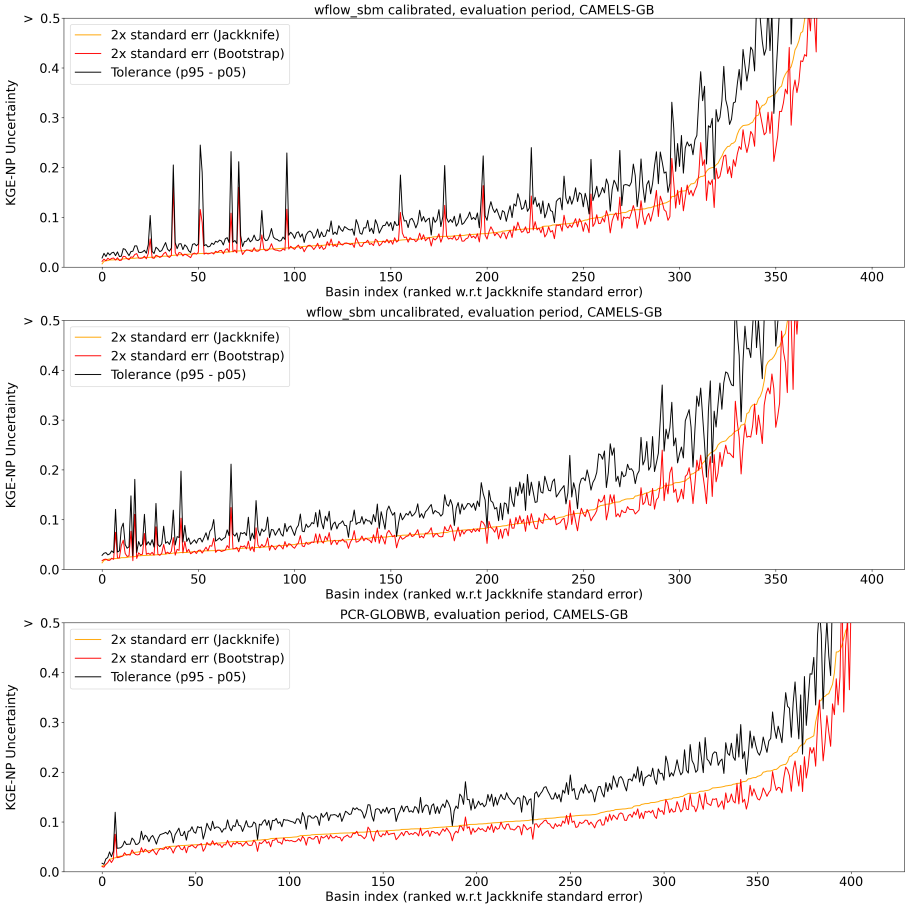


**Figure B1.** (A) Distributions of the temporal sampling uncertainty based on the KGE-NP objective function for the three model configurations. With the tolerance interval in black, the 2x the standard error of the jackknife method in orange and 2x the standard error of the bootstrap method in black. The horizontal axis is ranked with the respect to the jacknife standard error. This figure matches those of Clark et al. (2021)

for consistency.

https://doi.org/10.5194/egusphere-2023-1156
Preprint. Discussion started: 5 June 2023


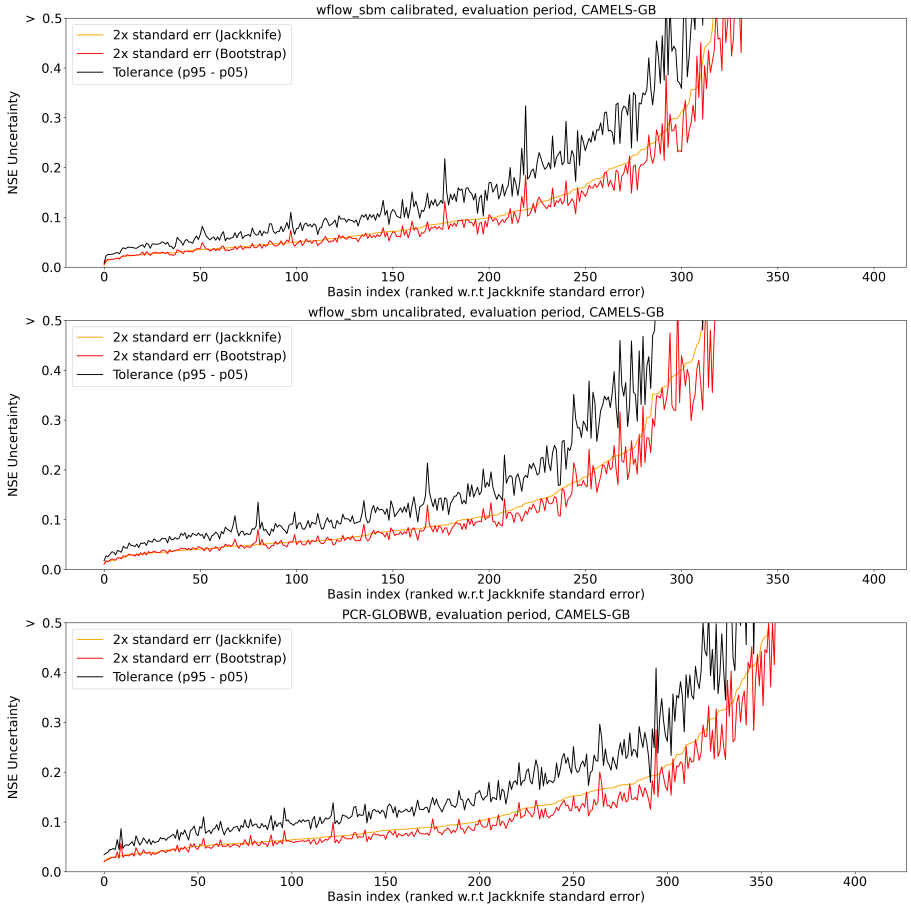


**Figure C1.** (A) Distributions of the temporal sampling uncertainty based on the NSE objective function for the three model configurations. With the tolerance interval in black, the 2x the standard error of the jackknife method in orange and 2x the standard error of the bootstrap method in black. The horizontal axis is ranked with the respect to the jacknife standard error. This figure matches those of Clark et al. (2021) for consistency.