On the importance of discharge observation uncertainty when interpreting hydrological model performance

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. For users of hydrological models, the suitability of models can be dependent upon how well their simulated outputs align with observed discharge. This study emphasizes the crucial role of factoring in discharge observation uncertainty when assessing the performance of hydrological models. We introduce an ad-hoc approach, implemented through the eWaterCycle platform, to evaluate the significance of differences in model performance while considering the uncertainty associated with discharge observations. The analysis of the results encompasses 299 catchments from the CAMELS-GB large-sample catchment dataset, addressing 3 distinct use cases that are of practical importance for model users. These use cases involve assessing the impact of additional calibration on model performance using discharge observations, conducting conventional model comparisons, and examining how the variations in discharge simulations resulting from model structural differences compare with the uncertainties inherent in discharge observations.

Our results, based on the 5th to 95th percentile range of observed flow, highlight the substantial influence of discharge observation uncertainty on the interpretation of model performance differences. Specifically, when comparing model performance before and after additional calibration, we find that in 98 out of 299 instances, the simulation differences fall within the bounds of discharge observation uncertainty. This underscores the inadequacy of neglecting discharge observation uncertainty during calibration and subsequent evaluation processes. Furthermore, in the model comparison use case, we identify numerous instances where observation uncertainty masks discernible differences in model performance, underscoring the necessity of accounting for this uncertainty in model selection procedures. While our assessment of model structural uncertainty generally indicates that structural differences often exceed observation uncertainty estimates, few exceptions do exist. The comparison of individual conceptual hydrological models indicates that there are no clear trends between model complexity and subsequent model simulations falling within the uncertainty bounds of discharge observations.

Based on these findings, we advocate for the integration of discharge observation uncertainty into the calibration process and also into the reporting of hydrological model performance as has been done in this study. This integration ensures more accurate, robust, and insightful assessments of model performance, thereby improving the reliability and applicability of hydrological modeling outcomes for model users.

1 Introduction

Many fields in geoscience rely on uncertain data to accurately estimate state and fluxes that support decision-making. Uncertain data in hydrology encompasses multiple sources that include direct measurements, proxy-based measurements, interpolation techniques, scaling processes, and data management practices (McMillan et al. (2018). 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); Beven et al. (2011); Beven and Smith (2015); Beven (2016); Beven and Lane (2022); Beven et al. (2022)). Data uncertainty can be distinguished into input data uncertainty (e.g. Kavetski et al. (2006a, b)) and evaluation data uncertainty (e.g McMillan et al. (2010).

Input data, primarily comprises meteorological variables such as precipitation and temperature. Other input data sources include static data, such as soil and topographic properties that are used to estimate model parameters. The inherent uncertainties in input datasets influence the model's simulation of states and fluxes (e.g. Balin et al. (2010); Bárdossy and Das (2008); Bárdossy et al. (2022); Bárdossy and Anwar (2023); McMillan et al. (2011); Beven (2021)). The uncertainty propagation from input to model output is also closely influenced by the model structure (Butts et al., 2004; Liu and Gupta, 2007; Zhou et al., 2022; Montanari and Di Baldassarre, 2013). The effects of uncertainty propagation have therefore been a focal point in literature, e.g. Beven (2006); Montanari and Toth (2007); Gupta and Govindaraju (2019).

Evaluation data uncertainty, the focus of this study, plays a pivotal role in determining the potential accuracy and robustness of hydrological models. This is the case for model calibration, a processes that involves fine-tuning model parameters to ensure that the model accurately and consistently reflects the observed historical behaviour of the hydrologic system. Typically this is based on discharge. When a model aims to replicate discharge values without including discharge observation uncertainty, the results are constrained to match a precise but potentially not accurate representation of the hydrological response (Vrugt et al., 2005). As a consequence, accurately calibrating the model becomes more challenging due to the demand of incorporating evaluation data uncertainty into the calibration process to minimize bias in model parameters (McMillan et al., 2010).

Multiple studies have demonstrated the importance of accounting for uncertainties in discharge observations. These mainly focus on hydrological model calibration (e.g. Beven and Binley (1992); Beven and Freer (2001); Beven and Smith (2015); McMillan et al. (2018); Beven and Lane (2019); Westerberg et al. (2020, 2022); McMillan et al. (2010); Coxon et al. (2015); Liu et al. (2009); Blazkova and Beven (2009)). In these studies multiple methodologies are used to quantify uncertainty estimates of discharge observations that are subsequently used for model calibration (overview in McMillan et al. (2012)).

Combined, all uncertainty sources (input data, evaluation data, model structure, model parameters, initial conditions) add to a concept in hydrological modelling commonly referred to as the equifinality concept (Beven and Freer (2001); Beven (2006); Montanari and Grossi (2008); Clark et al. (2008); Beven et al. (2011)). This concept is characterised by the circumstance of various model configurations yielding similar behavioural or acceptable results. Therefore, the recommendation is to account for all uncertainty sources simultaneously. An example of a method that includes all uncertainty sources during the parameter estimation process is the General Likelihood Uncertainty Estimation (GLUE; Beven and Freer (2001)) method. In practice, such methodologies are not always applied by model users although the difficulty of implementation can be dispelled (Pappenberger and Beven (2006)).

Hydrological model evaluation by model users is often solely based on discharge observations. The inherent uncertainties in this single source of observational data might obscure the model's ability to simulate actual discharge. Therefore, omitting data uncertainty during model evaluation negatively affects the interpretation of relative model simulation differences as these might fall within the uncertainty bounds of the observations.

Another challenging aspect of hydrological modelling is the large spatial and temporal variability of the hydrological system. Capturing the 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 catchment datasets contain hydro-meteorological time series, catchment boundaries and catchment attributes for a large number of catchments. The dataset is complemented with discharge observations at the catchment outlets and meteorological forcing datasets that include precipitation and temperature. The large-sample catchment datasets are collected using 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. More recently, Coxon et al. (2020) released the CAMELS-GB, that includes estimates of quantified discharge observation uncertainty. This dataset describes 671 catchments in Great Britain of which 503 gauging stations are complemented with quantified discharge observation uncertainty estimates (Coxon et al. (2015)). 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 catchment data triggered a wealth of research as discussed in the overview by Addor et al. (2020), including use 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 these datasets are that large-samples of catchments allow for the evaluation of the robustness of model performance (Andréassian et al. (2006); Gupta et al. (2014)). Identifying this robustness provides model users with valuable information on the presence or absence of consistency in the model results.

In this study, we assess the effect of omitting discharge observation uncertainty while interpreting model performance differences. Specifically, we focus on how this uncertainty influences model selection from the perspective of model users. Thereby, we highlight the importance of incorporating discharge observation uncertainty during model calibration and model evaluation efforts. To achieve this, we developed a generic method that is applicable for any geoscience field where model results are compared to uncertain observations. This method determines, based on the 5th to 95th percentile range of flow, if model simulation differences are significant in the context of discharge observation uncertainties. In this study, we highlight 3 use cases based on 8 hydrological models that encompass model refinement efforts, conventional model comparisons, and the influence of model structure uncertainty in light of discharge observation uncertainty. Furthermore, we assess the spatial consistency of model performance results using a large-sample catchment dataset and we assess the temporal consistency of model perfor-

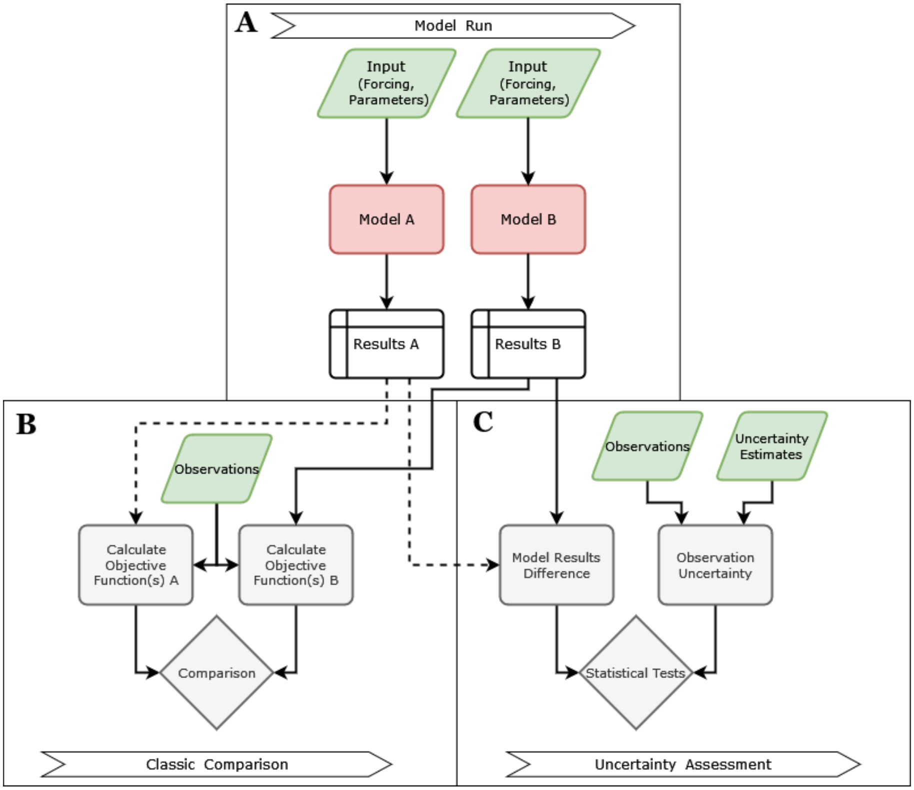

Figure 1. Graphical workflow of model experiments and analyses, with in green the model experiment inputs, in red the models, and in grey the analysis components. (a) the model runs of two models with inputs and outputs that result in simulation time series. (b) the conventional model comparison that compares objective functions based on simulation and observation time series. (c) the workflow of the proposed analysis that compares relative model simulation differences to discharge observation uncertainty estimates.

mance metrics by sub-sampling the observation and simulation pairs as demonstrated by Clark et al. (2021). By doing so, more informed conclusions can be drawn on model performance based on individual catchments or large-samples of catchments.

2 Methodology

A generic tooling is designed for assessing model simulations while considering the uncertainties inherent in evaluation data. First, the 3 use cases are presented, this is followed by the input data description, evaluation data description, and the discharge observation uncertainty estimates used to conduct the analyses. Next, we describe the employed hydrological models and model runs for calibration and evaluation. The methodology concludes with the explanation of the uncertainty based analyses.

In Figure 1, a graphical workflow is presented that provides an overview of the presented in the methodology. Figure 1a shows a typical model run with inputs and outputs, Figure 1b shows a conventional comparison of objective functions based on discharge observations and simulations, and Figure 1c describes the uncertainty analysis introduced in this study.

2.1 Use cases

We devised 3 use cases based on 8 hydrological models that exemplify how users of models, whom themselves are not the model developers, can interpret differences between model simulations in the context of discharge observation data uncertainty. The use cases are:

1. Model refinement in practice: This use case concerns additional model refinement by fine tuning an effective model parameter based on discharge observation post initial calibration. It highlights the value of relative gains in model performance when not considering discharge observation uncertainty in the calibration process.
2. Model comparison for model selection: Here, two distributed hydrological models are compared against the backdrop of uncertainties in discharge observations. This analysis aims to pinpoint scenarios where the disparities between model results are within the margin of error of the discharge observations.
3. Model selection under model structural uncertainty: This use case involves contrasting the uncertainty inherent in the model's structure, as seen across various hydrological models, with the uncertainty in discharge observations.

An additional analysis is performed that quantifies uncertainty in the model performance objective functions due to temporal sampling of the discharge simulation and observation pairs. This temporal sampling uncertainty is detailed in Section 2.5.3.

2.2 Data

2.2.1 Case study and catchment selection procedure

The CAMELS-GB large-sample catchment dataset (Coxon et al., 2020; Coxon, 2020) serves as the case study area of the use cases and contains data (hydro-meteorological time series, 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 hydrological models as the dataset can be easily extended in the future. A unique feature of the dataset is the availability of quantified discharge observation uncertainty estimates for the flow percentiles of 503 catchments (see Coxon et al. (2015)).

The use cases in this study employ hydrological models with a daily time step. This can cause temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. Therefore, these catchments are excluded through a selection procedure. This procedure calculates the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded. Of the 503 catchments with uncertainty estimates, 299 catchments are selected as the case study.

2.2.2 Meteorological forcing and pre-processing

In this study We use the same meteorological forcing that was used to create the CAMELS-GB meteorological time series 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)). The distributed hydrological models use gridded inputs and the conceptual hydrological models aggregated time series of meteorological variables that are readily available in the CAMELS-GB dataset.

2.2.3 Discharge observations and quantified uncertainty estimates

The discharge observations in the CAMELS-GB dataset were obtained from the UK National River Flow Archive and are daily values in cubic meters per second (m^3*s^{-1}). As is common with large-sample catchment datasets several catchments contain missing flow data in the time series. These missing values are not taken into account in the analyses of this study.

A unique aspect of the CAMELS-GB dataset is the inclusion of quantified discharge 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 discharge 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.

2.3 Hydrological models

A mixture of distributed physically process-based and lumped conceptual hydrological models is selected for the use cases, thereby showcasing the versatility of the analysis. The model refinement and model comparison use cases employ two physically process-based hydrological models: wflow_sbm (van Verseveld et al. (2022)) and PCR-GLOBWB (Sutanudjaja et al. (2018); Hoch et al. (2023)). The rationale behind selecting these models lies in their differing approaches to conceptualizing hydrological processes and their respective optimization routines. Despite these differences, both models are suitable for comparison to a certain degree. This comparability stems from their shared classification as distributed hydrological models, similar complexity, parameterization, and applicability at a spatial resolution of 1 km^2.

For the model structure use case, 6 conceptual hydrological models are sourced from the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT: Knoben et al. (2019); Trotter et al. (2022)). These specific models are selected to encompass a wide array of model structures. The selection is based on the number of model stores, the quantity of parameters, and differing process representations.

2.3.1 Distributed hydrological models

The wflow_sbm physically-based distributed hydrological model (van Verseveld et al. (2022)) originated from 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 in this study. The parameters (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)).

The 1 km^2 model version was aggregated from the finest available data scale (90 m). The hydraulic parameters related to the river network are upscaled using the method presented in Eilander et al. (2021). The parameter upscaling of the wflow_sbm model is based on the work by Imhoff et al. (2020) that uses point-scale (pedo)transfer-function. This method is similar to the multiscale parameter regionalization method (Samaniego et al. (2010)). Parameters are aggregated from the original data resolution with upscaling operators determined by a constant mean and standard deviation across various scales. Fluxes and states are checked for consistency during this process. See van van Verseveld et al. (2022) for further information.

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 determined by water demand. We employ 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.

2.3.2 Conceptual hydrological models

MARRMoT is a flexible modelling framework that houses an array of conceptual hydrological models Knoben et al. (2019); Trotter et al. (2022). It is particularly valued in research for assessing model structure uncertainty, as highlighted in Knoben et al. (2020). One of the key advantages of MARRMoT is that the conceptual models share a uniform numerical implementation. To achieve this, alterations were made to the original model codes. These alterations ensure a consistent basis for model structure comparisons, allowing for a precise evaluation of differences in hydrological simulations due to varying model structures. In this study, the hydrological models IHACRES, GR4J, VIC, XINANJIANG, HBV-96, and SMAR are selected. Table 1 provides an overview of the number of stores, number of parameters, and key references.

2.4 Model runs

The model runs that form the basis of the 3 use cases are performed as intended by the model developers. Meaning, this study employs calibration and or optimization methodologies as recommended by the model developers for model users. The calibrated parameters for the distributed hydrological models were obtained from the model developers. In the case of the conceptual hydrological models we follow the model run configuration of Knoben et al. (2020).

Table 1. Overview of the 6 selected conceptual hydrological models showing the model name, number of stores, number of parameters, and key references (adapted from Knoben et al. (2020)).

Original Model	Number of Stores	Number of Parameters	Key References
IHACRES	1	7	Ye et al. (1997); Croke and Jakeman (2004)
GR4J	2	4	Perrin et al. (2003); Santos et al. (2018)
VIC	3	10	Liang et al. (1994)
XINANJIANG	4	12	Jayawardena and Zhou (2000)
HBV-96	5	15	Lindström et al. (1997)
SMAR	6	8	Tan and O'Connor (1996)

2.4.1 PCR-GLOBWB model runs

The PCR-GLOBWB model does not require additional regional parameter optimization after deriving the parameter set, as this is typically not conducted by the model developers. However, this does not imply that the model would not benefit from additional optimization. The model does require a spin-up period 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 following 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.

2.4.2 Default and optimized wflow_sbm model runs

The wflow_sbm model is spun-up using the water year 2000 and additionally calibrated using discharge observations for the water years 2001-2007. Additional calibration is perfomered by optimizing a single parameter using the Kling-Gupta Efficiency Non-Parametric (KGE-NP) objective function (Pool et al. (2018)) based on discharge observations and simulations differences at the catchment outlet. This results in a single optimized parameter set per catchment. Imhoff et al. (2020) identified the horizontal conductivity fraction parameter (KsatHorFrac) as effective for single parameter value optimization. KsatHorFrac is an amplification factor of the vertical saturated conductivity that controls the lateral flow in the subsurface.

After calibration, 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. The calibration results of the wflow_sbm model are presented in Appendix A1.

2.4.3 Conceptual hydrological model runs

Similar to the other model runs, the conceptual hydrological model runs are spun-up using the water year 2000 and calibrated using the water years 2001-2007. The calibration method uses the Coviariance Matrix Adaptation Evolution Strategy (CMA-ES; Hansen et al. (2003); Hansen (2006); Hansen and Ostermeier (2001)). This method optimizes a single-objective function to find global parameter optimums based on non-separable data problems. A demonstration of the sensitivity of the calibration parameters is shown in Knoben et al. (2020). Following calibration based on the KGE-NP objective function, the water year 2008 is discarded and the models are evaluated based on the water years 2009-2015 using the same KGE-NP objective function.

2.4.4 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 model performance 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.

2.5 Analyses

2.5.1 Model evaluation

The hydrological model runs (calibration and evaluation) 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 -∞ 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.

2.5.2 Discharge observation uncertainty

The ad hoc discharge observation uncertainty based analysis of model performance differences consists of 3 parts. The first part divides the observation and simulation pairs into 3 flow categories similar to Coxon et al. (2015), namely low flow, average flow, and high flow conditions. The low flow category is based on the observed discharge values at the catchment outlet between the 5th and 25th percentile range, average flow on the 25th to 75th percentile range, and high flow on the 75th to 95th percentile range. Not all percentiles are included for the low and high flow categories due to limited data availability on quantified discharge observation uncertainty.

In the second part, illustrated in Figure 1c, the absolute difference between model simulations is calculated for each flow category and each catchment. This is exemplified in the form of a hydrograph in Figure 2a. The discharge observation uncertainty estimates of the CAMELS-GB dataset are processed by averaging the upper and lower bounds of uncertainty estimates per flow percentile (5, 25, 75, 95). This results in the orange and red dashed lines in Figure 2b. We then take the percentage of discharge observations basd on the average uncertainty estimates to convert the uncertainty percentages to discharge observation uncertainty time series in m^3*s^{-1} (green line).

The third part applies a dependent t-test using the time series in Figure 2c with a 0.05 significance level to determine if the observation uncertainty time series is greater than the model simulation difference time series.

This method is subject to certain limitations, particularly regarding the use of the discharge observation uncertainty estimates. Due to absence of data the upper and lower 5th percentiles of flow could not be included while these data points can be most important for users to determine fit-for-purpose of a model. In addition, it is preferred to use the rating curve uncertainty rather than the uncertainty bounds of flow percentiles. We accept these limitations as we promote the use of existing dataset to ensure community participation into implementing the suggested evaluation procedure in other studies.

2.5.3 Temporal sampling uncertainty

Another aspect of model performance evaluations that might misinform model users is the sensitivity of objective functions to the temporal sampling of time series. Temporal sampling uncertainty determines if the error distribution of simulation and observation pairs is heavily skewed. A few data pairs might have a disproportionate effect on the calculated objective functions that are used to determine model performance. The inclusion or exclusion of these data points due to the selection of the calibration and evaluation period, hence alters the consistency of model performance over time.

To quantify the temporal sampling uncertainty of the KGE-NP objective function, we applied the methodology of Clark et al. (2021). This method sub-samples the simulation and observation time series through bootstrapping and (Efron, 1979) and jacknife-after-bootstrap (Efron and Tibshirani, 1986) methods. The change in objective function due to the shuffling of the sub-samples allows for the calculation of the standard error and its tolerance interval. The tolerance intervals corresponding to each model instance are averaged and referred to as temporal sampling uncertainty. We extended the GUMBOOT software package Clark et al. (2021) by adding the KGE-NP metric for this study.

3 Results

In this section we first present an overview of the discharge-based model performance results for each of the 3 use cases. Next, we detail the spatial distributions of the maximum model performance difference. This is succeeded by the presentation of uncertainty estimates for discharge observations, categorized by flow. Subsequently, the discharge observation uncertainty based analyses of relative model performance is presented. The section ends with the temporal sampling uncertainty analysis results.

Appendix A.1 contains the calibration results of the wflow_sbm model and Appendix A.2 the Nash-Sutcliffe efficiency (NSE) based model performance results of all considered models.

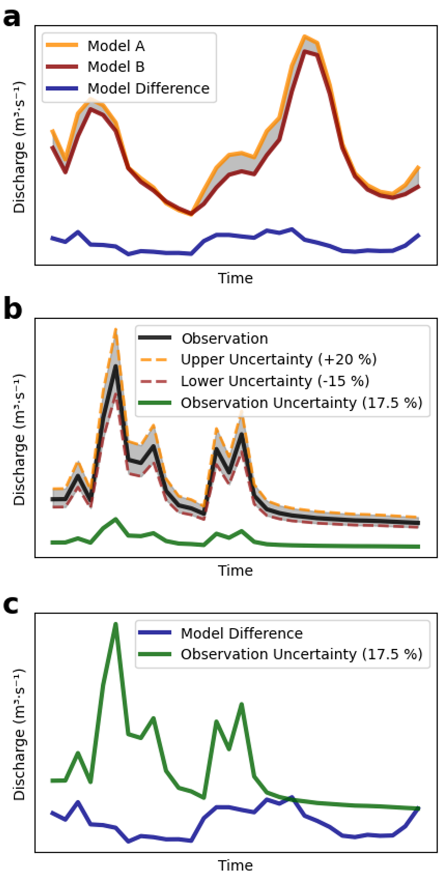

Figure 2. Example hydrographs of the discharge observation uncertainty analysis method. (a) calculation of the absolute difference (blue) between model simulations (red and orange). (b) calculation of streamflow observation uncertainty in m^3*s^{-1} (green). Dashed lines indicating upper and lower bounds expressed as percentages of observation uncertainty that are averaged and multiplied with the observations (black). (c) resulting time series, with, in blue, the absolute difference between model simulations and, in green, the averaged discharge observation uncertainty in m^3*s^{-1}.

3.1 Discharge based model performance

Model performance is assessed using discharge observation and simulations at 299 catchment outlets. The results are shown in Figure 3 as Cumulative Distribution Functions (CDFs) of the KGE-NP objective function. These results offer insight into the model's accuracy in simulating observed discharge.

The CDF of the model refinement use case in Figure 3a establishes that optimizing a single effective parameter leads to an improvement for approximately 65% of the catchments. The improvements remain modest as indicated by the median value of 0.64 KGE-NP for the default wflow_sbm model and 0.74 KGE-NP for the optimized wflow_sbm model. Larger model performance differences are found for the model comparison use case in Figure 3b. Here, the optimized wflow_sbm

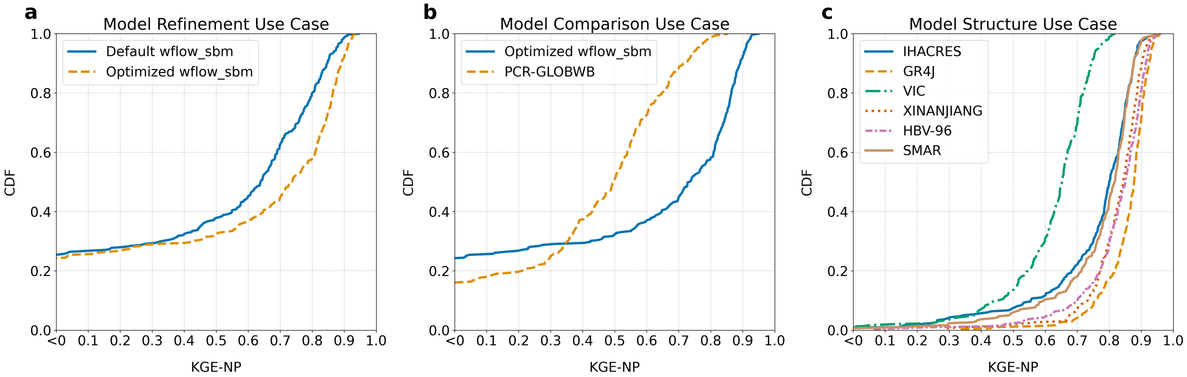

Figure 3. Cumulative distribution function (CDF) plots of the Kling-Gupta Efficiency non-parametric (KGE-NP) objective function, derived from discharge estimates and observations at 299 catchment outlets. (a) shows the CDF for the model refinement use case, optimizing the wflow_sbm hydrological model with a single parameter. (b) shows the CDF for the model comparison use case, comparing the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) demonstrates the CDF for the model structure use case, showcasing results from 6 conceptual hydrological models.

model performs better in 75% of the catchments than the PCR-GLOBWB model. Both models demonstrate poor results for approximately 25% of the evaluated catchments (<0.40 KGE-NP).

The results of the model structure use case are based on 6 conceptual hydrological models that only deviate in model structure (Figure 3c). From the spread in model results it is evident that the VIC model lags behind in performance compared to the other models. The IHACRES and SMAR models yield very similar results despite large structural differences. The XINANJIANG and HBV-96 models not only produce comparable outcomes but also share a more similar model structure. The GR4J model consistently outperforms the other models. The total model structure uncertainty, as expressed by the difference between the worst and best performing model's CDF is substantial, while the differences between models can be subtle. Median KGE-NP values for the models are as follows: VIC at 0.65, IHACRES at 0.80, SMAR at 0.82, XINANJIANG at 0.84, HBV-96 at 0.85, and GR4J at 0.88 KGE-NP.

Next, we consider the spatial distribution of the results, presented in Figure 4, based on the maximum KGE-NP difference between the models of each use case. Improvements after model refinement are indicated by the positive KGE-NP difference values in Figure 4a. These values are mainly present in the Northern and Southern parts of Great Britain. No clear spatial patterns are visible for the model comparison use case in Figure 4b, demonstrating high spatial variability in performance when comparing the wflow_sbm and PCR-GLOBWB distributed hydrological models. More spatially consistent differences are found for the model structure use case in Figure 4c. Here, the largest differences are present in the Northern and Southern parts of Great Britain.

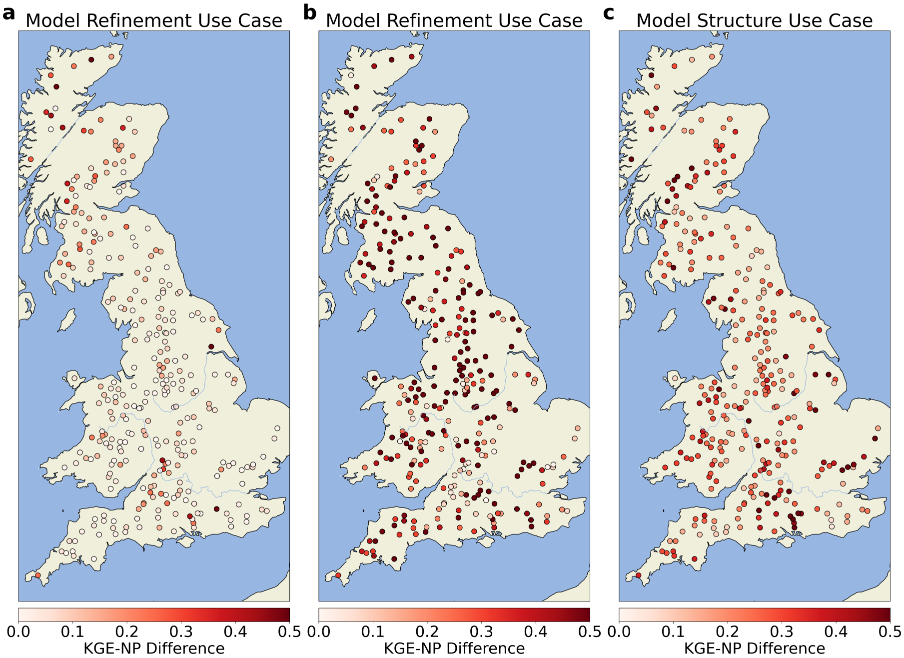

Figure 4. Spatial distribution of the absolute Kling-Gupta Efficiency non-parametric (KGE-NP) objective function difference between the worst and best model's performance per catchment and use case. (a) shows the model refinement use case based on the default and optimized wflow_sbm hydrological models. (b) shows the model comparison results based on the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) shows the model structure use case results based on the worst and best model performances of the 6 conceptual hydrological models.

3.2 Discharge observation uncertainty estimates

The discharge observation uncertainty estimates consider the 5th to 95th percentile range of flow. These estimates are categorized into 3 flow conditions and are presented in Figure 5. In the box plot for low flow category, we observe a wide interquartile range, shown by the spread of the box. This indicates a higher variability in discharge observation uncertainty percentages. The median value, represented by the line within the box, is at the 20% uncertainty mark. The presence of many outliers above the box indicate occasional large deviations from the median value. For the average flow category, the range of values is narrower than for the low flow category with a median value of 15%. The lowest median value is found for the high flow category at 12%. It is important to mention that the uncertainty is expected to be considerably higher if the underlying data would contain the upper 5th percentiles of flow for this category.

3.3 Use Cases

The discharge simulation difference time series of two models is expressed in cubic meters per second and compared to discharge observation uncertainty time series in cubic meters per second. This is done by using a t-test to determine if the

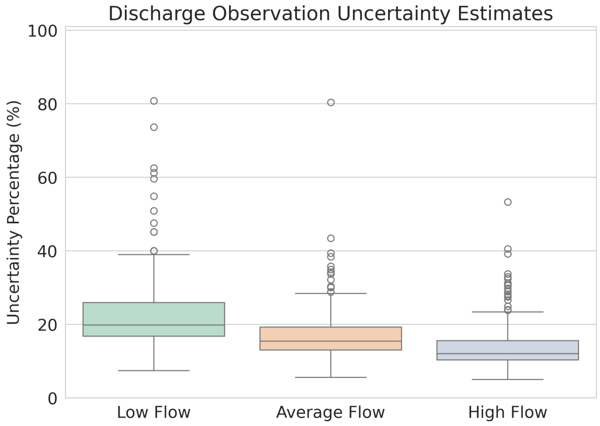

Figure 5. Discharge observation uncertainty estimates of 299 catchment outlets based on the work of Coxon et al. (2015) expressed as uncertainty percentages per flow category. (a) shows the low flow category uncertainty estimates based on the 5th to 25th flow percentiles. (b) presents the 25th to 75th percentile average flow category. (c) shows the high flow discharge observation uncertainty estimates of the 75th to 95th flow percentiles.

Table 2. Overview of the number of instances per flow category were discharge observation uncertainty exceeds the simulation differences based on 299 catchments. Results are based on dependent t-test's with a significance level of 0.05.

Use Case	**Models**	**Flow Category**	**Discharge Obs. Uncertainty > Model Sim. Difference**	**Total Instances**
Model Refinement	wflow_sbm Default & Optimized	Low	98	299
		Average	98	
		High	115	
Model Comparison	wflow_sbm Optimized & PCR-GLOBWB	Low	5	299
		Average	4	
		High	3	
Model Structure	6 Conceptual Hydrological Models	Low	1	299
		Average	0	
		High	0	

simulation differences are larger than the discharge observation uncertainty estimates. The instances where this is the case are reported in Table 2 for the 3 use cases.

3.3.1 Model refinement

The model refinement use case results in Table 2 show that approximately one third of the considered catchments contain instances of simulation differences between the wflow_sbm default and wflow_sbm optimized models that are statistically smaller than the discharge observation uncertainty estimates. This demonstrates the importance of incorporating (discharge) observation uncertainty when performing model refinement, especially when based on a large-sample catchment dataset. This consideration should be part of the calibration and subsequent evaluation process. In addition, the results indicate that when discharge observation uncertainty is not considered, it is difficult to draw conclusions on whether the model performs better after refinement. Overall, the results affirm the importance of incorporating discharge observation uncertainty in the optimization routine of wflow_sbm model.

3.3.2 Model comparison

For the model comparison use case (Table 2), there is a lower frequency of instances where discharge observation uncertainty surpasses differences in discharge simulations. The comparison between the optimized wflow_sbm model and the PCR-GLOBWB model reveals that in 5 catchments for low flow, 4 for average flow, and 3 for high flow categories, simulation differences exceed discharge uncertainty estimates. These findings suggest that the interpretation of model performance is not significantly affected by the ad-hoc addition of discharge observation uncertainty. However, catchments demonstrating the impact of observation uncertainty warrant carefull examination.

3.3.3 Model structure

The analysis of model structure uncertainty in the context of discharge observation uncertainty reveals that only a single instance of the low flow category contains discharge observation uncertainty that exceed the simulation difference between all 6 conceptual hydrological models (Table 2). This establishes that based on the selected models the model structure uncertainty, expressed as the difference in discharge simulations, is larger than the discharge observation uncertainty estimates for this dataset. However, the investigation into the differences between the individual models yields several insights based on the results in Figure 6.

The VIC model results, characterized by its relatively lower performance, contain only a few instances where discharge observation uncertainty exceeds simulation differences, making it identifiable as the lesser performing model. In contrast, the IHACRES and SMAR models exhibit a high level of simulation agreement, demonstrated by a large number of instances in Figure 6c. This, despite significant differences in their complexity and structural design. Namely, IHACRES is a single store hydrological model and SMAR is a 6 store hydrological model that accounts for soil moisture in a separate store. This alignment of simulation results between models with varying complexities highlights the nuanced influence of structural differences on simulation outcomes. The HBV-96 and XINANJIANG models that most closely resemble each other based on the number of stores, process descriptions, and parameters contain low number of instance, allowing the identification of the better performing model.

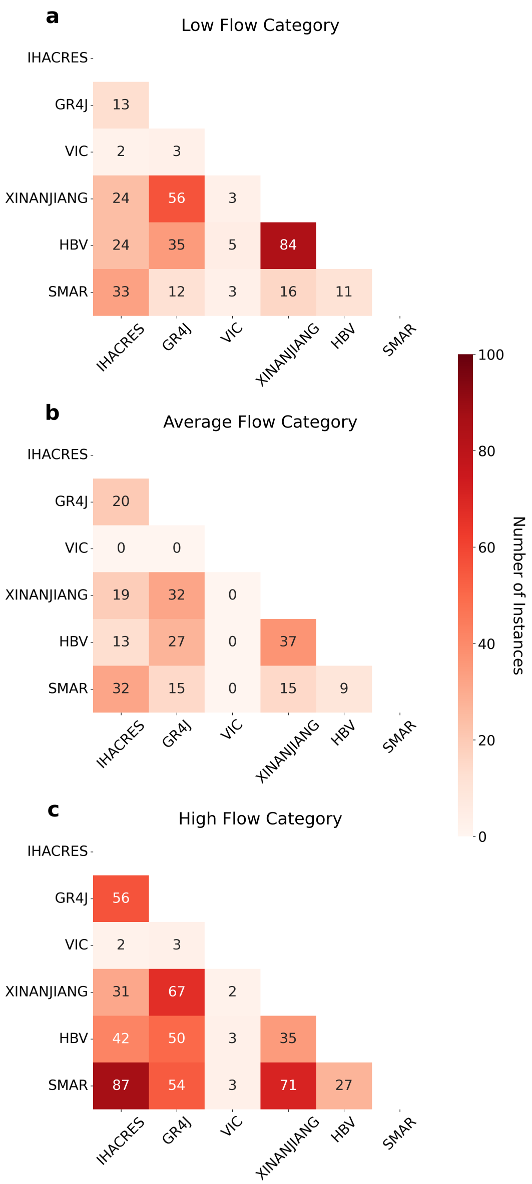

Figure 6. Heat map of the 6 conceptual hydrological models showing for each model combination the number of instances (n=299) that discharge observation uncertainty exceeds simulation differences. (a) number of instances for the low flow category, with in white low values and in red high values. (b) number of instances for the average flow category. (c) number of instances for the high flow category.

Next, we examine the results across the individual flow categories. The low flow category (Figure 6a) and the average flow category (Figure 6b) show similar trends, though with a lower number of instances for the average flow category with a lower number of instances for the average flow category. The high flow category (Figure 6c) is characterized by a more frequent 345 occurrence of discharge observation uncertainty surpassing simulation differences. This is especially evident between the IHACRES and SMAR models. The variability in structural design and parameterization among different hydrological models leads to notable differences in their outputs, underscoring the importance of selecting the appropriate model by including discharge observation uncertainty in the calibration and evaluation process.

3.4 Temporal sampling uncertainty

The temporal sampling uncertainty of the KGE-NP objective function is defined as the tolerance interval of the standard error of the objective function due to sub-sampling of the simulation and observation pairs. This analysis provides insights into the temporal reliability and interpretability of hydrological model performances. Analysis of results from the 6 conceptual hydrological models, as shown in Figure 7b, reveals a pattern consistent with the model performance depicted in Figure 3c. Specifically, the VIC model displays the highest KGE-NP uncertainty across all catchments, indicating its variability and the challenges in using this model current setup for accurate predictions in different hydrological contexts.

The IHACRES and SMAR models, along with GR4J, XINANJIANG, and HBV-96, show similar levels of KGE-uncertainty. This consistency across models with varying complexities suggests that KGE-NP uncertainty is influenced not only by the model design but also by hydrological conditions and data quality. Uncertainty values range widely, from about 0.1 KGE-NP to over 0.6, indicating significant variability in temporal robustness of results (Figure 7b).

When comparing the average KGE-NP objective function uncertainty with the KGE-NP differences between individual models, it becomes clear that uncertainty often overshadows the differences between models. This is particularly the case in comparisons between GR4J - HBV-96, XINANJIANG - HBV-96, and SMAR - IHACRES. These findings imply that the inherent uncertainty in the objective functions may limit the ability to distinguish between model performances, complicating efforts to identify the most fit-for-purpose model based on this metric alone. This underscores the need for a more nuanced approach to model evaluation that considers not only objective function metrics but also other contextual factors and additional performance measures, ensuring more robust and reliable model selection processes.

4 Discussion

We introduced an ad hoc method that highlights the importance of including discharge observation uncertainty when evaluating hydrological models. Discharge observation uncertainty is frequently overlooked by model users, leading to potential misinterpretations of relative model performance. Our findings emphasize the significant impact of discharge observation uncertainty on model performance interpretation.

We acknowledge that observation uncertainty is not the only source of uncertainty as there are uncertainties in model inputs, model structure, parameter sets, and initial or boundary conditions (e.g. Renard et al. (2010); Dobler et al. (2012); Hattermann et al. (2018); Moges et al. (2021)). Therefore the proposed generic tooling does not replace a full uncertainty analysis of modelling chains that also accounts for the impact of input uncertainties (Beven and Freer (2001); Pappenberger and Beven (2006); Beven (2006)). It rather assists model users with interpreting relative model performance and highlights the importance

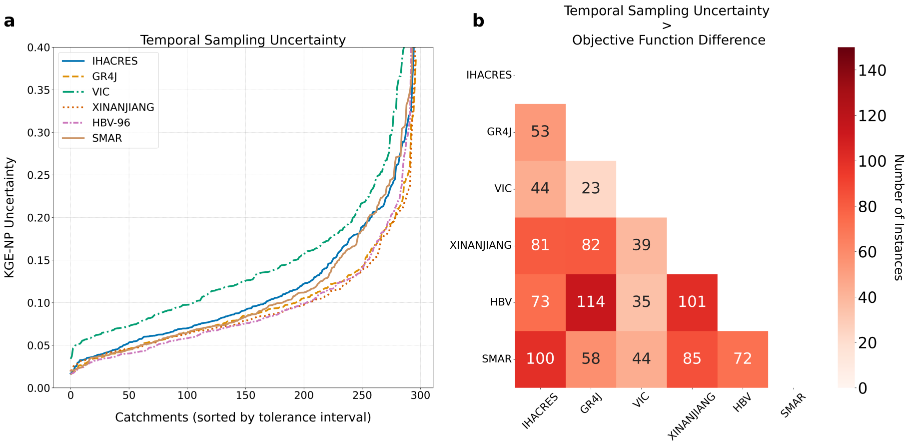

Figure 7. (a) Temporal objective function sampling uncertainty based on 6 conceptual hydrological models expressed as the average tolerance interval of the standard error due to sub-sampling. With on the horizontal axis the sorted values per catchment and on the vertical axis the KGE-NP objective function uncertainty. (b) Heat map of the 6 conceptual hydrological models showing for each model combination the number of instances (n=299) that the average objective function uncertainty exceeds the objective functions differences of model combinations. With in white low values and in red high values indicating the number of instances.

of conducting a full uncertainty analysis. Our study therefore only constitutes just a fraction of a broader challenge, in which input uncertainty plays a substantial role as has been demonstrated in Bárdossy and Anwar (2023).

4.1 Performance interpretation under discharge observation uncertainty

Our analysis demonstrates that regionally optimizing the wflow_sbm hydrological model often results in only marginal improvements in model performance (Figure 3a). Although any improvement is beneficial, the findings suggest that discerning the superior model variant becomes challenging without factoring in the uncertainty of discharge observations during the calibration process. This is evident in 98 instances for low and average categories of flow and in 118 instances of the high flow category (Table 2). The number of instances is expected to further increase when including flows of the lower and upper 5th percentiles of flow are included. The adoption of an ad hoc measure, as introduced in this study, provides a practical though limited method for improving the interpretability of relative model results. Therefore, we recommend the integration of discharge observation uncertainty into both the model calibration and evaluation procedures, aligning with the consensus in literature.

When comparing different hydrological models, we find that the uncertainty of discharge observations slightly masks the differences in relative model performance as shown by the 3 to 5 instances per flow category in Figure 3b and Table 2. Similarly to the model comparison use case, the model structure use case indicates that structural uncertainties overshadow the effects of discharge observation uncertainty. However, the comparison of individual models in Figure 6 shows many instances of discharge observation uncertainty exceeding model performance differences. For instance, the IHACRES and SMAR models, despite their structural differences, demonstrate a high level of simulation agreement (Figure 3c) and subsequent difficulty in discerning model performance differences in light of discharge observation uncertainty. In contrast, the VIC and XINANJIANG models, which have similar structures, display for two-thirds of the catchment simulation differences within the uncertainty bounds of the discharge observations. This underlines the complex interplay of model structures and subsequent performance, especially when contrasted with discharge observation uncertainty.

4.2 Temporal robustness of model performance

Model performance can be heavily influenced by a few data points in the time series on which model performance is based (Clark et al. (2021)). This can results in biased model performance interpretations depending on the selected time period for calibration and evaluation. When models are sensitive to certain data points this can be due to a lack of adequate process descriptions in the considered models. In addition, this might also indicate the presence of disinformative events and model invalidation sites where the runoff coefficient exceeds a value of 1 (Beven and Smith (2015); Beven (2023); Beven and Lane (2022); Beven et al. (2022)) or the presence of atypical data (e.g. Thébault et al. (2023)).

Models ought to demonstrate adequate performance across the entire time series and this should be accurately represented in the performance outcomes. The assessment of temporal sampling uncertainty does not imply that this is should not be the case, it rather points towards in the model simulation and observation pairs that are worth investigating. These instances can serve as indicators that suggest areas where models may require further scrutiny and improvements. Knowing the temporal sampling uncertainty is relevant for model users as it provides information on the consistency of the model performance over time that is necessary to determine the fit-for-purpose of a model. Therefore, it is recommended to include alternative estimators better suited for skewed performance data in the reporting of model performance (e.g. Lamontagne et al. (2020); Shabestanipour et al. (2023); Towler et al. (2023)).

4.3 Practical implications for model users

The method introduced in this study is purposely designed to be as generic and straightforward as possible to increase the potential for adoption in future studies. It can be applied to any hydrological state or flux where observation time series include uncertainty estimates. In addition, we recommend for the routine reporting of evaluation data uncertainties as well as the temporal sampling uncertainty of objective functions. This would not only yield a clearer understanding of the relevance of differences between model outcomes but also aid in identifying samples that require cautious interpretation. This reporting, however, does not replace model benchmarks that include full uncertainty analyses (e.g. Lane et al. (2019)), but enhances the interpretability of model performance in its absence.

For model users, this approach offers a pragmatic way to understand the implications of uncertainty in their model selection processes. While our method facilitates a clearer understanding of where and how uncertainties affect relative model performance differences, it should be viewed as a complementary step rather than a replacement for a thorough uncertainty analysis.

4.4 Limitations

The study presented faces several practical limitations. First, the exclusion of the lower and upper 5th percentiles of flow from the analysis introduces a constraint on the uncertainty assessment, overlooking critical flow conditions that are often of significant interest in hydrological studies. This exclusion limits the ability to fully understand model performance under a complete range of hydrological conditions. Second, the reliance on uncertainty bounds rather than direct uncertainty estimates from rating curves, due to their absence in the CAMELS-GB dataset, poses another limitation. By using broad uncertainty bounds instead of precise estimates derived from rating curves, the analysis may not capture the true variability and uncertainty inherent in the discharge observations. Last, the study's focuses solely on evaluating model performance primarily through discharge simulations, without delving into the reasons behind good or poor model performance as this is outside of the scope of the study.

5 Conclusions

This study assesses the importance of including discharge observation uncertainty and temporal sampling uncertainty of objective functions in hydrological model performance evaluations based on a large-sample catchment dataset. This is done by statistical testing that determines if the difference in discharge simulations between two hydrological models is larger or smaller than the discharge observation uncertainty estimates. To support this analysis flow categories are created between the 5th and 95th percentile range of observed flow and 3 use cases are devised.

In the model refinement use case a substantial 100 out of 299 catchment instances showed discharge simulation differences between the default and optimized wflow_sbm models that were within the uncertainty bounds of discharge observations. This emphasizes the need for integrating discharge observation uncertainty in the calibration process for model refinement. As a result it is difficult to discern if the optimization of the model leads to improved simulations of actual discharge. For the model comparison use case, we found that depending on the model combinations a large fraction of catchments showed discharge observation uncertainty exceeding simulation differences. Thereby suggesting careful consideration of this uncertainty in model performance evaluations. The model structure uncertainty use case that is based on 6 conceptual hydrological models indicated only a few instances of discharge observation uncertainty exceeding simulation differences. Indicating that model structure uncertainty, expressed as discharge simulation differences, often exceeds discharge observation uncertainty. Comparison of the six individual hydrological models showed no clear relation between model complexity and model performance.

Our study underscores the necessity of integrating discharge observation uncertainty and temporal sampling uncertainty into hydrological model evaluations to ensure accurate, reliable, and meaningful assessments of model performance. Implementing our proposed methodology in reporting practices is expected to improve the robustness of hydrological model result interpretation, aiding in more informed model selection and refinement decisions by model users.

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].

References

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.

Balin, D., Lee, H., and Rode, M.: Is point uncertain rainfall likely to have a great impact on distributed complex hydrological modeling?, Water Resources Research, 46, https://doi.org/10.1029/2009WR007848, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2009WR007848, 2010.

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.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrological Sciences Journal, 61, 1652–1665, https://doi.org/10.1080/02626667.2015.1031761, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2015.1031761, 2016.

Beven, K.: An epistemically uncertain walk through the rather fuzzy subject of observation and model uncertainties1, Hydrological Processes, 35, e14 012, https://doi.org/10.1002/hyp.14012, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.14012, 2021.

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

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, https://doi.org/10.1002/hyp.3360060305, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.3360060305, 1992.

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. and Lane, S.: Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach, in: Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, edited by Beisbart, C. and Saam, N. J., Simulation Foundations, Methods and Applications, pp. 145–171, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-70766-2_6, 2019.

Beven, K. and Lane, S.: On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose, Hydrological Processes, 36, e14 704, https://doi.org/10.1002/hyp.14704, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.14704, 2022.

Beven, K. and Smith, P.: Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models, Journal of Hydrologic Engineering, 20, A4014 010, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000991, publisher: American Society of Civil Engineers, 2015.

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.

Beven, K., Lane, S., Page, T., Kretzschmar, A., Hankin, B., Smith, P., and Chappell, N.: On (in)validating environmental models. 2. Implementation of a Turing-like test to modelling hydrological processes, Hydrological Processes, 36, e14 703, https://doi.org/10.1002/hyp.14703, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.14703, 2022.

Blazkova, S. and Beven, K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, Water Resources Research, 45, https://doi.org/10.1029/2007WR006726, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2007WR006726, 2009.

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, Journal of Hydrology, 298, 242–266, https://doi.org/10.1016/j.jhydrol.2004.03.042, 2004.

Bárdossy, A. and Anwar, F.: Why do our rainfall–runoff models keep underestimating the peak flows?, Hydrology and Earth System Sciences, 27, 1987–2000, https://doi.org/10.5194/hess-27-1987-2023, publisher: Copernicus GmbH, 2023.

Bárdossy, A. and Das, T.: Influence of rainfall observation network on model calibration and application, Hydrology and Earth System Sciences, 12, 77–89, https://doi.org/10.5194/hess-12-77-2008, publisher: Copernicus GmbH, 2008.

Bárdossy, A., Kilsby, C., Birkinshaw, S., Wang, N., and Anwar, F.: Is Precipitation Responsible for the Most Hydrological Model Uncertainty?, Frontiers in Water, 4, https://www.frontiersin.org/journals/water/articles/10.3389/frwa.2022.836554, 2022.

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/10.1029/2007WR006735, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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/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.

Croke, B. F. W. and Jakeman, A. J.: A catchment moisture deficit module for the IHACRES rainfall-runoff model, Environmental Modelling & Software, 19, 1–5, https://doi.org/10.1016/j.envsoft.2003.09.001, 2004.

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.

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.

Eilander, D., van Verseveld, W., Yamazaki, D., Weerts, A., Winsemius, H. C., and Ward, P. J.: A hydrography upscaling method for scale-invariant parametrization of distributed hydrological models, Hydrology and Earth System Sciences, 25, 5287–5313, https://doi.org/10.5194/hess-25-5287-2021, publisher: Copernicus GmbH, 2021.

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.

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, A. and Govindaraju, R. S.: Propagation of structural uncertainty in watershed hydrologic models, Journal of Hydrology, 575, 66–81, https://doi.org/10.1016/j.jhydrol.2019.05.026, 2019.

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.

Hansen, N.: The CMA Evolution Strategy: A Comparing Review, in: Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms, edited by Lozano, J. A., Larrañaga, P., Inza, I., and Bengoetxea, E., Studies in Fuzziness and Soft Computing, pp. 75–102, Springer, Berlin, Heidelberg, https://doi.org/10.1007/3-540-32494-1_4, 2006.

Hansen, N. and Ostermeier, A.: Completely Derandomized Self-Adaptation in Evolution Strategies, Evolutionary Computation, 9, 159–195, https://doi.org/10.1162/106365601750190398, 2001.

Hansen, N., Müller, S. D., and Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evolutionary Computation, 11, 1–18, https://doi.org/10.1162/106365603321828970, 2003.

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., 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.

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.

Jayawardena, A. W. and Zhou, M. C.: A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model, Journal of Hydrology, 227, 93–113, https://doi.org/10.1016/S0022-1694(99)00173-0, 2000.

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.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42, https://doi.org/10.1029/2005WR004368, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004368, 2006a.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, Water Resources Research, 42, https://doi.org/10.1029/2005WR004376, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004376, 2006b.

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., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, publisher: Copernicus GmbH, 2019.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, Water Resources Research, 56, e2019WR025 975, https://doi.org/10.1029/2019WR025975, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR025975, 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.

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/10.1029/2020WR027101, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR027101, 2020.

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.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research: Atmospheres, 99, 14 415–14 428, https://doi.org/10.1029/94JD00483, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/94JD00483, 1994.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, https://doi.org/10.1029/2006WR005756, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2006WR005756, 2007.

Liu, Y., Freer, J., Beven, K., and Matgen, P.: Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, Journal of Hydrology, 367, 93–103, https://doi.org/10.1016/j.jhydrol.2009.01.016, 2009.

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., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, Journal of Hydrology, 400, 83–94, https://doi.org/10.1016/j.jhydrol.2011.01.026, 2011.

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.

McMillan, H. K., Westerberg, I. K., and Krueger, T.: Hydrological data uncertainty and its implications, WIREs Water, 5, e1319, https://doi.org/10.1002/wat2.1319, 2018.

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.

Montanari, A. and Di Baldassarre, G.: Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty, Advances in Water Resources, 51, 498–504, https://doi.org/10.1016/j.advwatres.2012.09.007, 2013.

Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: A statistical approach, Water Resources Research, 44, https://doi.org/10.1029/2008WR006897, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2008WR006897, 2008.

Montanari, A. and Toth, E.: Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?, Water Resources Research, 43, https://doi.org/10.1029/2006WR005184, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2006WR005184, 2007.

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.

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.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

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.

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

Santos, L., Thirel, G., and Perrin, C.: Continuous state-space representation of a bucket-type rainfall-runoff model: a case study with the GR4 model using state-space GR4 (version 1.0), Geoscientific Model Development, 11, 1591–1605, https://doi.org/10.5194/gmd-11-1591-2018, publisher: Copernicus GmbH, 2018.

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., and Lamontagne, J. R.: Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation, Water Resources Research, 59, e2022WR032 201, https://doi.org/10.1029/2022WR032201, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022WR032201, 2023.

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.

Tan, B. Q. and O'Connor, K. M.: Application of an empirical infiltration equation in the SMAR conceptual model, Journal of Hydrology, 185, 275–295, https://doi.org/10.1016/0022-1694(95)02993-1, 1996.

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.

Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., and Delaigue, O.: Impact of suspicious streamflow data on the efficiency and parameter estimates of rainfall–runoff models, Hydrological Sciences Journal, 68, 1627–1647, https://doi.org/10.1080/02626667.2023.2234893, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2023.2234893, 2023.

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://doi.org/10.5194/hess-27-1809-2023, publisher: Copernicus GmbH, 2023.

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., and Peel, M. C.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: an object-oriented implementation of 47 established hydrological models for improved speed and readability, Geoscientific Model Development, 15, 6359–6369, https://doi.org/10.5194/gmd-15-6359-2022, publisher: Copernicus GmbH, 2022.

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.

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water Resources Research, 41, https://doi.org/10.1029/2004WR003059, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2004WR003059, 2005.

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.

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., and Seibert, J.: Hydrological model calibration with uncertain discharge data, Hydrological Sciences Journal, 67, 2441–2456, https://doi.org/10.1080/02626667.2020.1735638, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2020.1735638, 2022.

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.

Ye, W., Bates, B., Viney, N., Sivapalan, M., and Jakeman, A.: Performance of Conceptual Rainfall-Runoff Models in Low-Yielding Ephemeral Catchments, Water Resources Research - WATER RESOUR RES, 33, 153–166, https://doi.org/10.1029/96WR02840, 1997.

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.

Zhou, L., Liu, P., Gui, Z., Zhang, X., Liu, W., Cheng, L., and Xia, J.: Diagnosing structural deficiencies of a hydrological model by time-varying parameters, Journal of Hydrology, 605, 127 305, https://doi.org/10.1016/j.jhydrol.2021.127305, 2022.

1 A.1 wflow_sbm calibration

A1 A.2 NSE based model performance results

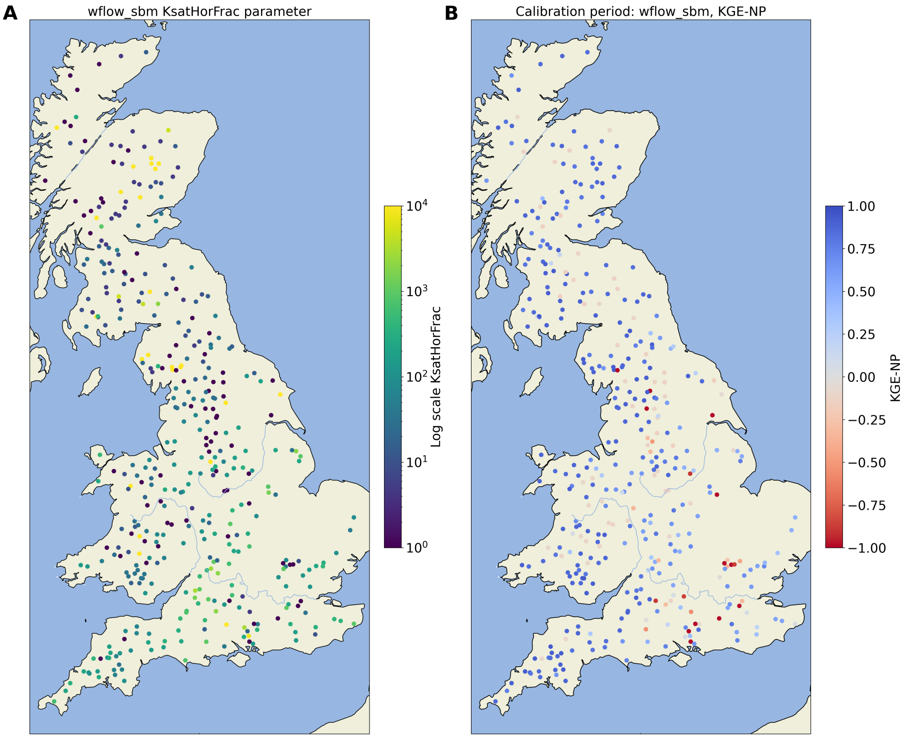

Figure A1. (A) Spatial distribution of the best performing KsatHorFrac calibration parameter of the wflow_sbm model based on additional calibration on streamflow observations. (B) Spatial distribution of the KGE-NP objective function based on the calibration period of the wflow_sbm model.

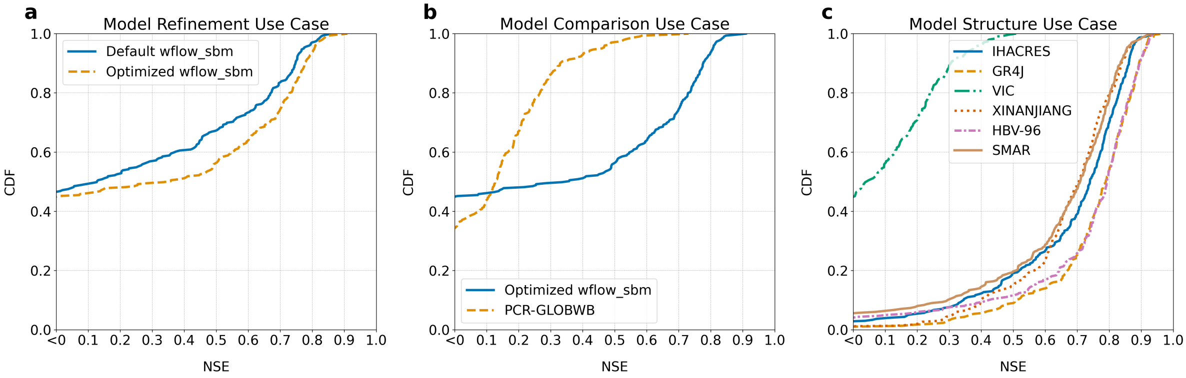

Figure B1. Cumulative distribution function (CDF) plots of the Nash-Sutcliffe Efficiency (NSE) objective function, derived from discharge estimates and observations at 299 catchment outlets. (a) shows the CDF for the model refinement use case, optimizing the wflow_sbm hydrological model with a single parameter. (b) shows the CDF for the model comparison use case, comparing the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) demonstrates the CDF for the model structure use case, showcasing results from six conceptual hydrological models.