



Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model

Jerom P. M. Aerts¹, Rolf W. Hut¹, Nick C. van de Giesen¹, Niels Drost², Willem J. van Verseveld³, Albrecht H. Weerts^{4,5}, and Pieter Hazenberg⁶

¹Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, the Netherlands

²The Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, the Netherlands

³Catchment and Urban Hydrology, Department of Inland Water Systems, Deltares, P.O. Box 177, 2600MH Delft, the Netherlands

⁴Operational Water Management, Department of Inland Water Systems, Deltares, P.O. Box 177, 2600MH Delft, the Netherlands

⁵Hydrology and Quantitative Water Management Group, Wageningen University and Research, P.O. Box 47, 6700AA Wageningen, the Netherlands

⁶Applied Research Center, Florida International University, FL 33174 Miami, United States of America

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

Received: 25 November 2021 – Discussion started: 1 December 2021

Revised: 17 June 2022 – Accepted: 31 July 2022 – Published: 29 August 2022

Abstract. Distributed hydrological modelling moves into the realm of hyper-resolution modelling. This results in a plethora of scaling-related challenges that remain unsolved. To the user, in light of model result interpretation, finer-resolution output might imply an increase in understanding of the complex interplay of heterogeneity within the hydrological system. Here we investigate spatial scaling in the form of varying spatial resolution by evaluating the streamflow estimates of the distributed wflow_sbm hydrological model based on 454 basins from the large-sample CAMELS data set. Model instances are derived at three spatial resolutions, namely 3 km, 1 km, and 200 m. The results show that a finer spatial resolution does not necessarily lead to better streamflow estimates at the basin outlet. Statistical testing of the objective function distributions (Kling–Gupta efficiency (KGE) score) of the three model instances resulted in only a statistical difference between the 3 km and 200 m streamflow estimates. However, an assessment of sampling uncertainty shows high uncertainties surrounding the KGE score throughout the domain. This makes the conclusion based on the statistical testing inconclusive. The results do indicate strong locality in the differences between model instances expressed by differences in KGE scores of on average 0.22

with values larger than 0.5. The results of this study open up research paths that can investigate the changes in flux and state partitioning due to spatial scaling. This will help to further understand the challenges that need to be resolved for hyper-resolution hydrological modelling.

1 Introduction

Hydrological model development follows competing model philosophies (Hrachowitz and Clark, 2017). From one end of the spectrum to the other these include high-resolution, small-scale, process-resolving distributed models and spatially lumped conceptual models. All model structures developed within the competing philosophies have their own limitations and advantages. There are overlapping challenges that all models face including parameter estimation and the representation of spatial heterogeneity (Clark et al., 2017).

The parameter identifiability problem stems from the inability to obtain unique and realistic parameters at the modelling scale due to structural model deficiencies and applied calibration techniques (Sorooshian and Gupta, 1983). Multiple studies have extensively researched calibration tech-

niques to overcome the parameter identifiability problem (e.g. Sorooshian and Gupta, 1983; Vrugt et al., 2002; Guse et al., 2020). The identifiability problem is emphasized in distributed modelling, the focus of this study, by the limitation of parameter measurements not being compatible with the modelling scale (Grayson et al., 1992). This results in the need for transferring parameters in space and time. Multiple studies have looked into parameter transferability (e.g. Finnerty et al., 1997; Haddeland et al., 2006; Wagener and Wheeler, 2006). Melsen et al. (2016) discussed that the inadequacy of transferring parameters in space and time may indicate a lack in spatial heterogeneity and temporal variability representation in the models. Methods such as the multi-scale parameter regionalization (MPR) technique (Samaniego et al., 2010) emerged to increase the transferability of parameters from the data resolution to the hydrological model resolution. Imhoff et al. (2020) used a different method that applied pedotransfer functions (PTFs) to derive parameters at the finest available data resolution and upscale them to the various spatial model resolutions of the wflow_sbm hydrological model.

The effects of spatial heterogeneity has been studied at a catchment scale using the representative elementary watershed (REW) theory developed by Wood et al. (1988); Reggiani et al. (1998, 1999); Reggiani and Schellekens (2003). In Wood et al. (1988) the basin was divided in sub-basins and then aggregated to basin level to study scaling behaviour. This type of research is still very relevant as in recent years, the hydrologic modelling community has surpassed the REW scale threshold of 1 km² with the move towards so called hyper-resolution modelling. The discussion following this move revealed that in addition to the many benefits, e.g. applicability for stakeholders, there are multiple challenges to address (Wood et al., 2011; Beven and Cloke, 2012; Bierkens et al., 2015). These challenges include scaling issues (Gupta et al., 1986; Blöschl and Sivapalan, 1995) such as (1) the need to explicitly model processes that are parameterized at coarser resolutions, (2) lateral connections between compartments of the hydrological system that are averaged out or ignored at coarser resolutions, and (3) an increase in uncertainty due to lacking process and parameter knowledge as a result of insufficient data quality at finer resolutions (Bierkens et al., 2015).

The scaling issues arise when the (often unconscious) assumption is made that a hydrological model used at various spatial and temporal resolutions should estimate similar states and fluxes independent of scale. A utopian model has scale-invariant model parameterization and hydrological process descriptions. The development of scale-invariant hydrological models is, however, very challenging as most hydrological processes do not scale in a linear manner (e.g. Bras, 2015; Rouholahnejad Freund et al., 2020). Instead, processes at one length scale influence those at other scales (Horritt and Bates, 2001).

Due to the complex nature of scaling issues and a shifting distributed modelling climate towards hyper-resolution modelling, it is important to continuously assess the effects of scaling. Without investigating what this move entails, the hydrological modelling community risks communication problems with the users of model results. To the user, in the case of spatial model resolution, the increase in the level of detail in model output might imply an increase in understanding of the complex interplay of heterogeneity within the hydrological system. We can only determine this by continuously assessing how models behave under various spatial (and temporal) resolutions.

Multiple studies have looked into spatial scaling effects by varying spatial model resolution. Booij (2005) found that increasing the spatial resolution of a semi-lumped HBV hydrological model only marginally increased model performance based on streamflow estimates. The coupled ParFlow-CLM model was evaluated with various grid-cell sizes by Shrestha et al. (2015) and they found, among other effects, that soil moisture estimates were dependent on grid-cell size. Sutanudjaja et al. (2018) introduced the transition from 30 to 5 arcmin grid-cell size simulations of the distributed PCR-GLOBWB model. Results showed a general increase in model performance compared to streamflow observations. However, regional scaling issues were present. In some of the basins model performance was lower at a finer spatial model resolution. This study made it apparent that a large sample of hydrologically diverse basins should be considered when investigating the effects of varying spatial resolution. To our knowledge there are no studies that have looked into varying spatial resolution within the hyper-resolution realm on a large sample of basins. This will be the focus of this research.

The distributed conceptual wflow simple bucket model (wflow_sbm) (Schellekens et al., 2020) utilizes high-resolution data sets to derive model instances globally at varying spatial resolution. Parameter estimates are based on the work of Imhoff et al. (2020) to ensure consistency across scales. Remotely sensed soil and land cover data sets are sources for estimating parameters through PTFs (e.g. Brakensiek et al., 1984; Cosby et al., 1984). The PTFs are a collection of predictive functions, so-called super parameters (Tonkin and Doherty, 2005), derived at point scale that estimate soil parameters where underlying data are scarce. For most wflow_sbm model parameters, a priori parameters are available. This is not (yet) the case for the horizontal conductivity fraction (KsatHorFrac) parameter, making it a logical parameter for calibration as it is also one of the more sensitive parameters in the model (Imhoff et al., 2020). The flexible setup of wflow_sbm can be used to assess spatial scaling issues due to quasi-scale invariant parameters whilst maintaining similar hydrological process descriptions across scales. This setup includes the recent improvements by Eilander et al. (2021), who developed a scale-invariant method for upscaling river networks (one of the suggested causes of

the inconsistent streamflow estimates across scales as shown in Imhoff et al., 2020).

In this study we quantify the effects of varying spatial resolution on the wflow_sbm streamflow estimates for a large sample of hydrological diverse basins in the CAMELS data set (Newman et al., 2015; Addor et al., 2017). By conducting this research on a large sample of basins, we can assess the results on consistency and locality. The assessment is conducted by creating three model instances at varying spatial resolutions for each basin: a 3 km, 1 km, and 200 m spatial grid resolution. These instances cover a broad range of large- and small-scale dynamics, for example, snow accumulation at the mountain range (> 1 km) and the mountain ridge scale (< 1 km) (e.g. Houze Jr., 2012; Mott et al., 2018; Vionnet et al., 2021), or closing in on the hillslope scale (< 100 m) (e.g. Tromp-van Meerveld and McDonnell, 2006; Fan et al., 2019). The parameters for the wflow_sbm model instances are estimated at the highest available data resolution and aggregated to the modelling grid using the upscale rules as defined in Imhoff et al. (2020).

Our hypothesis is that the differences in streamflow estimates at various spatial resolutions will be small due to the parameters being quasi-scale invariant and hydrological process descriptions in the model remaining the same across spatial scales. We will reject this hypothesis when the results show significantly different streamflow estimates across the studied resolutions. Additionally, we will showcase how the eWaterCycle platform (Hut et al., 2021) can be utilized for computational intensive large-sample modelling studies.

2 Methodology

2.1 Input data

2.1.1 The CAMELS data set

The CAMELS data set is a collection of hydrologically relevant data on 671 basins located in the contiguous United States (CONUS) (Newman et al., 2015). The basins were selected based on a minimum amount of human influence on the hydrological system, e.g. the absence of large reservoirs. The data set includes 20 years of continuous streamflow records from 1990 to 2009 from the United States Geological Survey (USGS). The CAMELS data set covers a hydrologically and hydro-climatologically diverse selection of basins. The sample-size, hydrological diversity of basins, and common use of this data set in other hydrological modelling studies (e.g. Knoben et al., 2020; Gauch et al., 2021) are the reasons for selecting this case study area.

Of the 671 basins, we ran 567 basins successfully for each of the three model instances (i.e. 3 km, 1 km and 200 m resolution). Reasons for excluding basins in our analysis are missing streamflow observations (7 basins) or errors during parameter estimation (97 basins). Parameter estimation er-

rors occurred mainly during drainage network delineation, when either the basin outlet consisted of a single grid cell that results in a model coding error or inconsistencies occurred in the local drainage direction layer. When a single model instance of the three model instances failed, the basin was excluded from further analysis. Figure 1 shows the locations of the included and excluded basins as well as the reason for exclusion.

2.1.2 Streamflow observations

The USGS streamflow observation records were downloaded to match our model simulation period from 1996 to 2016. The data are resampled to daily data and the units were converted to $\text{m}^3 \text{s}^{-1}$. We ensured consistency in time zones between the observation and model simulations by matching the USGS streamflow data with the UTC time zone. The tooling used for downloading, resampling, unit conversion, and shifting of time zones might be of interest to others in the hydrological community and is available in the GitHub repository (https://github.com/jeromaerts/eWaterCycle_example_notebooks, last access: 8 June 2022).

2.1.3 Meteorological input and pre-processing

The meteorological input requirements of the wflow_sbm model are precipitation, temperature, and potential evapotranspiration. Precipitation data were obtained from the Multi-Source Weighted-Ensemble Precipitation (MSWEP) Version 2.1 (Beck et al., 2019). The data set was constructed using bias-corrected gauge, satellite, and reanalyses data. The data are available at 0.1 degrees spatial (11 km) and 3 h temporal resolution for the period 1979–2017. The temperature variable was obtained from the ERA5 reanalyses data set (Hersbach et al., 2020). The data are available at 0.25° (31 km) spatial and 1 h temporal resolution. In addition to temperature, we used ERA5 variables to calculate potential evapotranspiration using the De Bruin method (Bruin et al., 2016).

We conducted a preliminary analysis for six basins in which we compared model simulations based on streamflow estimates that use ERA5 precipitation to those that use MSWEP precipitation. Results indicated that the use of ERA5 precipitation did not produce desirable streamflow estimates compared to MSWEP precipitation. Switching to the MSWEP precipitation product improved streamflow estimates throughout the case study area. Figure A1 and Table A1 in Appendix A contain the results of this analysis. Noticeably, some of the basins are very sensitive to changes due to different forcing data sets as shown by the streamflow-based objective functions.

The meteorological input is pre-processed within the eWaterCycle platform using the Earth system model evaluation tool (ESMValTool) Version 2.0 (Righi et al., 2020; Weigel et al., 2021). Before further processing, the data are aggre-

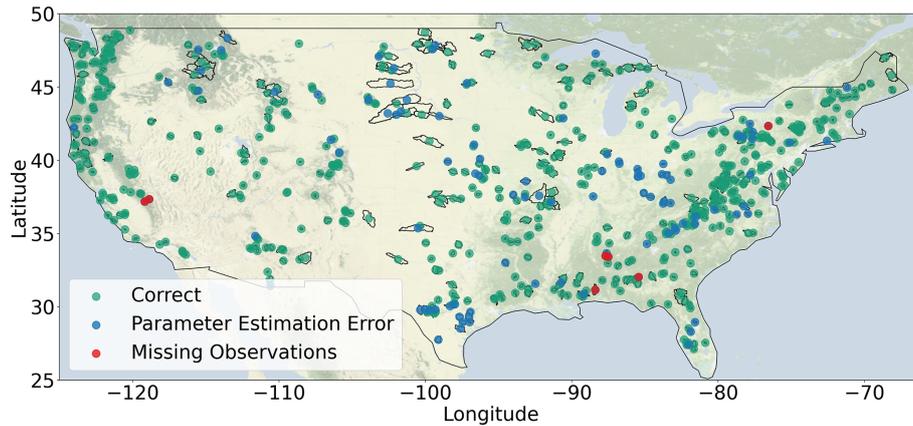


Figure 1. Basin locations of the CAMELS data set. The included basins are marked in green; blue shows the excluded basins due to parameter estimation errors and red shows the excluded basins due to missing streamflow observations. Basemap made with Natural Earth.

gated to daily values. The precipitation variable is disaggregated to the modelling grid using the second-order conservative method to ensure consistency of the total volume of precipitation across spatial scales. The temperature variable is disaggregated with the environmental lapse rate and the digital elevation model (DEM) used by the hydrological model. The variables required by the De Bruin method (Bruin et al., 2016) are disaggregated using the (bi)linear method and are subsequently used to calculate potential evapotranspiration. The code used for pre-processing is included in the Jupyter Notebooks made available with this paper (<https://doi.org/10.5281/zenodo.5724512>).

2.1.4 Parameter estimation from external data sources

The parameter sets used in this study were derived using the hydroMT software package (Eilander and Boisgontier, 2021). The data sources for deriving parameter sets are open-source global data sets. These include topography, surface water, land cover and land use, soil, meteorology, and river gauge data. The PTF to estimate soil properties is based on Brakensiek et al. (1984). In 25 of the 567 basins, lakes and/or reservoirs were included in the model parameters given a threshold area of 1 and 10 km², respectively.

An overview of the data and references are provided in Table 1.

2.2 Model experiment setup

2.2.1 The wflow_sbm model (v.2020.1.2)

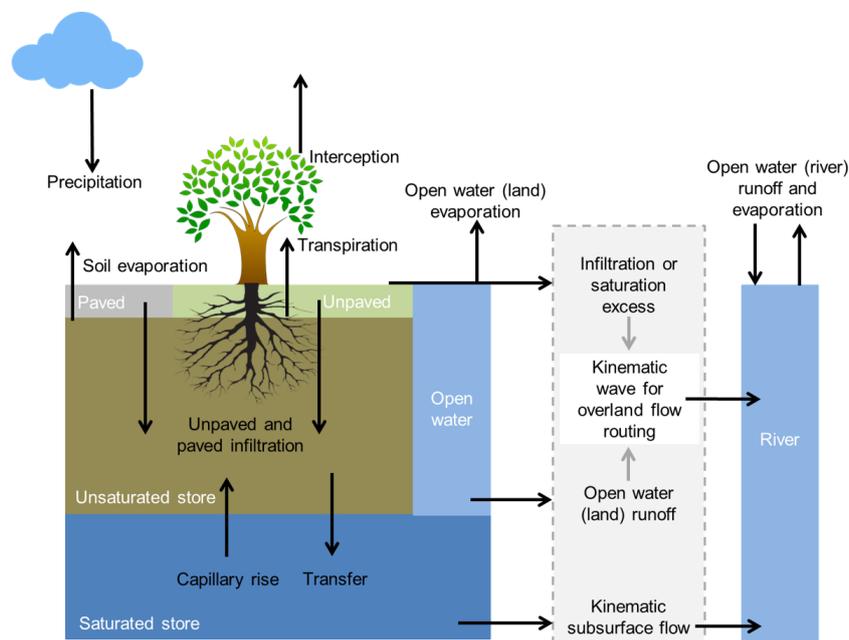
The wflow_sbm model is available as part of the wflow open-source modelling framework (Schellekens et al., 2020), which is based on PCRaster (Karszenberg et al., 2010) and Python. Figure 2 shows the different processes and fluxes that are part of the wflow_sbm hydrological concept. The soil part of the wflow_sbm model is largely based on the Topog_SBM model (Vertessy and Elsenbeer, 1999), which regards the soil

as a “bucket” with a saturated and unsaturated store. This model was developed for simulating small-scale hydrology. For channel, overland and lateral subsurface flow, a kinematic wave approach is used, similar to TOPKAPI (Benning, 1994; Ciarapica and Todini, 2002), G2G (Bell et al., 2007), 1K-DHM (Tanaka and Tachikawa, 2015), and Topog_SBM (Vertessy and Elsenbeer, 1999). The wflow_sbm model has a simplified physical basis with parameters that represent physical characteristics, leading to (theoretically) an easy linkage of the parameters to actual physical properties. The Topog_SBM model is mainly used to simulate fast runoff processes during discrete storm events in small basins (< 10 km²) (evapotranspiration losses are ignored). Since evapotranspiration losses and capillary rise were added to wflow_sbm, the derived wflow_sbm approach can be applied to a wider variety of basins. The main differences of wflow_sbm with Topog_SBM include

- the addition of evapotranspiration and interception losses using the Gash model (Gash, 1979) on daily time steps or a modified Rutter model on subdaily time steps (Rutter et al., 1971, 1975);
- the addition of a root water uptake reduction function (Feddes et al., 1978);
- the addition of capillary rise;
- the addition of glacier and snow build-up and melting processes;
- wflow_sbm that routes water over an eight direction (D8) network, instead of the element network based on contour lines and trajectories, used by Topog_SBM;
- the option to divide the soil column into any number of different layers;

Table 1. Overview of data sources for parameter estimation with categories, references, and version.

Data set	Category	Reference	Version
Merit Hydro	Topography	Yamazaki et al. (2019)	1.0
GRAND (hydro_reservoirs)	Surface water	Lehner et al. (2011)	1.0
hydroLAKES (hydro_lakes)	Surface water	Messenger et al. (2016)	1.0
Randolph Glacier Inventory	Surface water	Pfeffer et al. (2014)	6.0
CHELSA	Meteorology	Karger et al. (2017)	1.2
Köppen–Geiger	Meteorology	Kottek et al. (2006)	2017
VITO	Land use & land cover	Buchhorn et al. (2020)	v2.0.2
Modis LAI	Land use & land cover	Myneni et al. (2015)	MCD15A3H V006
SoilGrids	Soil	Hengl et al. (2017)	2017

**Figure 2.** Overview of the different processes and fluxes in the wflow_sbm model (Schellekens et al., 2020).

- vertical transfer of water that is controlled by the saturated hydraulic conductivity at the water table or bottom of a layer, the relative saturation of the layer, and a power coefficient depending on the soil texture (Brooks and Corey, 1964).

2.2.2 Model runs and calibration

We derived three model instances at varying spatial model resolutions that cover a 3 km, 1 km, and 200 m grid. While for most parameters of the wflow_sbm model a priori estimates can be derived from external sources, a single non-distributed parameter needs to be calibrated for each basin: the saturated horizontal conductivity often expressed as a fraction (KsatHorFrac) of the vertical conductivity. This parameter cannot be derived from external data sources because it compensates for anisotropy, unrepresentative point measurements of the saturated vertical conductivity, and model

resolution (Schellekens et al., 2020). A sensitivity analysis conducted by the model developers concluded that KsatHorFrac is the most effective parameter when it comes to calibration based on streamflow estimates, also briefly discussed in Imhoff et al. (2020). Increasing the value of this parameter leads to an increased base flow component and reduced peak flow and flashiness.

We calibrated the models to match model setups of those used by the users of the hydrological model. The model instances are calibrated using the modified Kling–Gupta efficiency score (KGE) (Gupta et al., 2009; Kling et al., 2012) by comparing the simulated streamflow estimates with the streamflow observations at the basin outlet. Twenty-one runs are evaluated based on an interval of KsatHorFrac values ranging between 1 and 1000. The best-performing model run and its corresponding KsatHorFrac parameter are selected for further analyses during the evaluation period. The model calibration is conducted from 1997 to 2005 and the model

evaluation from 2007 to 2016. The years 1996 and 2006 are regarded as spin-up years and are not included in the analysis.

2.3 Benchmark selection

To select basins with reasonably good model performance, we applied a statistical benchmark to beat. The use of a benchmark allows for better interpretation of objective function-based results (Garrick et al., 1978; Pappenberger et al., 2015; Schaeffli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018; Knoben et al., 2020). We adopt, in part, the same methodology for statistical benchmark creation as Knoben et al. (2020). The benchmark is created by calculating the mean and median of streamflow observations per calendar day of the 10-year evaluation period and the results are compared to streamflow predictions from the model. The KGEs are calculated for the observed streamflow versus these mean and median values. The mean is expected to better represent larger basins with more stable flow regimes whilst the median better covers the flashiness of smaller headwater basins. This benchmark serves as a lower boundary for the model predictions: if the model, for any of the three resolutions has a lower KGE than either the median or mean flow benchmark, it is considered to be unsuitable for this study and removed from further analyses.

2.4 Analysis of results

2.4.1 Objective function and sampling uncertainty

We use the modified KGE metric for the analysis of results. Ideal model performance has a KGE score of 1 and a KGE score of -0.41 is equal to taking the mean flow as a benchmark (Knoben et al., 2019c). The lowest value of the three components determines the KGE score. The KGE score and its components are as follows:

$$\text{KGE}' = 1 - \sqrt{(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}, \quad (1)$$

$$\beta = \left(\frac{\mu_s}{\mu_o} \right), \quad (2)$$

$$\gamma = \left(\frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \right), \quad (3)$$

where r is the Pearson correlation, β the mean bias, and γ the variability bias. The mean and standard deviation of the simulations is denoted as μ_s and σ_s , respectively and μ_o and σ_o are the mean and standard deviation of the observations.

We quantify the sampling uncertainty of the KGE score for the selected basins based on the statistical benchmark following the methodology of Clark et al. (2021). This applies the non-overlapping bootstrap method (Efron and Tibshirani, 1986) to calculate tolerance intervals and jackknife-after-bootstrap method (Efron, 1992) for the standard error calculation of those tolerance intervals. This method applies

bootstrap and jackknife methods to estimate the standard errors and tolerance interval of KGE uncertainty. The tolerance interval is defined as the difference between the 5th and 95th percentiles. We ran the gumboot analysis package (Clark et al., 2021) with a sample size of 500 during the evaluation period to calculate the results.

2.4.2 Comparison of streamflow estimates

To provide more context for the results in terms of general model performance, we compared the streamflow estimates from wflow_sbm to those of the study by Knoben et al. (2020). The latter ran 36 conceptual models on the CAMELS data set using the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) Version 1.0 (Knoben et al., 2019a, b). First, we calculated the mean of the 36 models for each basin. Next, we ensured a match between the basins under investigation by both studies. Due to differences in time period, forcing, and numerical solvers, the results cannot be compared directly to those of this study. It does however provide context for the results.

The inter-model (instance) comparison of the streamflow estimates in this study is assessed using a cumulative distribution function (CDF). We applied the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1933) to assess whether the differences between the KGE score distributions of the model instances are statistically relevant. This allows the acceptance or rejection of the hypothesis stating that the differences in streamflow estimates at various spatial resolutions will be small.

2.5 eWaterCycle platform

This research was conducted within the eWaterCycle platform (Hut et al., 2021). By design, eWaterCycle follows the FAIR (Findable, Accessible, Interoperable, and Reusable) principles of data science (Wilkinson et al., 2018) and allows high-level communication with models regardless of programming language through the Basic Model Interface (BMI; Hutton et al., 2020). This study showcases the way eWaterCycle handles the setup of extensive modelling studies. A Jupyter Notebook with the model experiments of this study is provided in the GitHub repository. Since notebooks are not ideal for long-running experiments on high-performance computing (HPC) machines, we exported the notebooks to regular python code which we ran directly on an HPC. The calibration and evaluation procedures totalled 41.025 model runs on the Dutch national supercomputer Cartesius hosted by Surf.

3 Results

The results in this section are based on the modified KGE 2012 objective function applied to the streamflow estimates at the basin outlet. The Nash–Sutcliffe efficiency (NSE)

(Nash and Sutcliffe, 1970) results are available in the repository (<https://doi.org/10.5281/zenodo.5724576>).

3.1 Calibration period results

3.1.1 The effect of calibration on streamflow estimates

To illustrate how model calibration affects the streamflow estimates of each model instance, we first show the calibration curves of a single basin (ID:14301000). We selected a basin with moderate performance and only show the last year of calibration to avoid presentation bias. Figure 3a–c shows the calibration curves (yellow to red) for each of the instances that were generated by tuning the horizontal conductivity (i.e. *KsatHorFrac*) parameter. The parameter values range from 1 to 1000 and are a single value for each basin. For visualization purposes, not all calibration interval values are shown in the figure. The results depict the effect that increasing the *KsatHorFrac* values has on the hydrograph: base flow increases while peak flow reduces. In addition, large *KsatHorFrac* values reduce the flashiness of the streamflow estimates as is visible in Fig. 3b and c in the second week of November 2005. It is worth noting that the selection of best calibration parameter values is strongly dependent on the chosen objective function as the NSE score would be more favourable towards flashiness and less towards base flow, for example. As shown in Fig. 3d, the streamflow estimates of the model instance are similar (KGE score 0.58–0.66) while the parameter values deviate (*KsatHorFrac* 125–1000). No strong apparent trends for *KsatHorFrac* values in relation to model resolution or geographic location were found after calibration.

Figure 4 shows the CDFs of the KGE score distribution based on the best-performing calibration model run of the three model instances. Starting with the modified KGE score distributions in Fig. 4a, we find that the model instance distributions are very similar for the entire domain, especially the 1 km and 200 m instances. The 3 km instance has lower scores for 60 % of the distributions between 0.2 and 0.8 CDF. Approximately 18 % of the distributions is lower than a KGE score of -0.41 which corresponds to taking the mean flow. All three model instances show a similar Pearson *r* correlation CDF (Fig. 4b) and gamma variability ratio (Fig. 4c). The largest differences are visible in the beta bias component (Fig. 4d) that shows similar bias for the 1 km and 200 m instances with 60 % of the distributions being lower than a value of 1 and 40 % being higher. The highest agreement is in the lower and upper 5 % of the distributions. The 3 km instance only agrees with the 1 km and 200 m instances in the upper 10 % of the distribution where 70 % of the 3 km distribution is lower than a value of 1 and 30 % is higher. The bias term of the KGE score has the largest weight in determining the shape of the KGE score CDF as is shown by 20 % of the distributions with bias values larger than 2.0.

3.2 Evaluation period results

3.2.1 Benchmark selection

The statistical benchmark is applied to determine which basins contain the streamflow estimates of the model instances that are deemed adequate for further analyses. The statistical benchmark is based on the best-performing type of climatology per calendar day, either mean or median, during the model evaluation period. Figure 5a spatially shows which type of calendar-day benchmark is performing best per basin. Of the 567 simulated basins, the results of 454 basins exceed the benchmark for each model instance. This is the case for all basins in the Midwest of the United States. Poor performance in comparison to the benchmark is mainly present in the Southwest. Based on the KGE scores, 83 % of the benchmark is favourable towards using the climatological mean and 27 % towards the median. An overview is provided in Fig. 5c and the distribution of the benchmark KGE scores is shown in Fig. 5d. The distribution ranges from a -0.62 to 0.71 KGE score and is skewed towards values lower than 0.0 with a mean of 0.02 and median of -0.02 .

The statistical benchmarks (mean and median during the evaluation period) are plotted as CDFs based on the KGE score and its three individual components in Fig. 6. For most of the KGE distributions in Fig. 6a, the mean benchmark outperforms the median with the upper 10 % being the exception. The Pearson correlation coefficient (Fig. 6b) is slightly higher for the median at 70 % of the distributions. In the case of the gamma variability ratio (Fig. 6c), the bottom 38 % is lower for the median than the mean, and the upper 55 % is slightly higher for the median than the mean. The determining component of the KGE score is the beta component shown in Fig. 6d. As expected, the mean benchmark is, for the most part, close to a value of 1, meaning that it is close to the mean of the observations. Of interest are the points of the median that are not close to 1, i.e. the lower 10 % and the upper 1 % of the distributions. These basins have flow regimes that greatly differ per year compared to the climatology. Considering the mean statistical benchmark of basins with a beta bias score lower than < 0.75 and higher than 1.75, we find that the model instances outperform the statistical benchmark in 35 of the 44 basins.

3.2.2 The effect of spatial resolution on streamflow estimates

As with the calibration period, the same three example basins are used to illustrate the differences in streamflow estimates between model instances for the evaluation period. Only the last year of the evaluation period is shown in Fig. 7.

In the case of poor performance (Fig. 7a), it may occur that the model instances are overestimating streamflow during peak flow. The best-performing model instance has the smallest peak-flow estimates which is the coarsest spatial res-

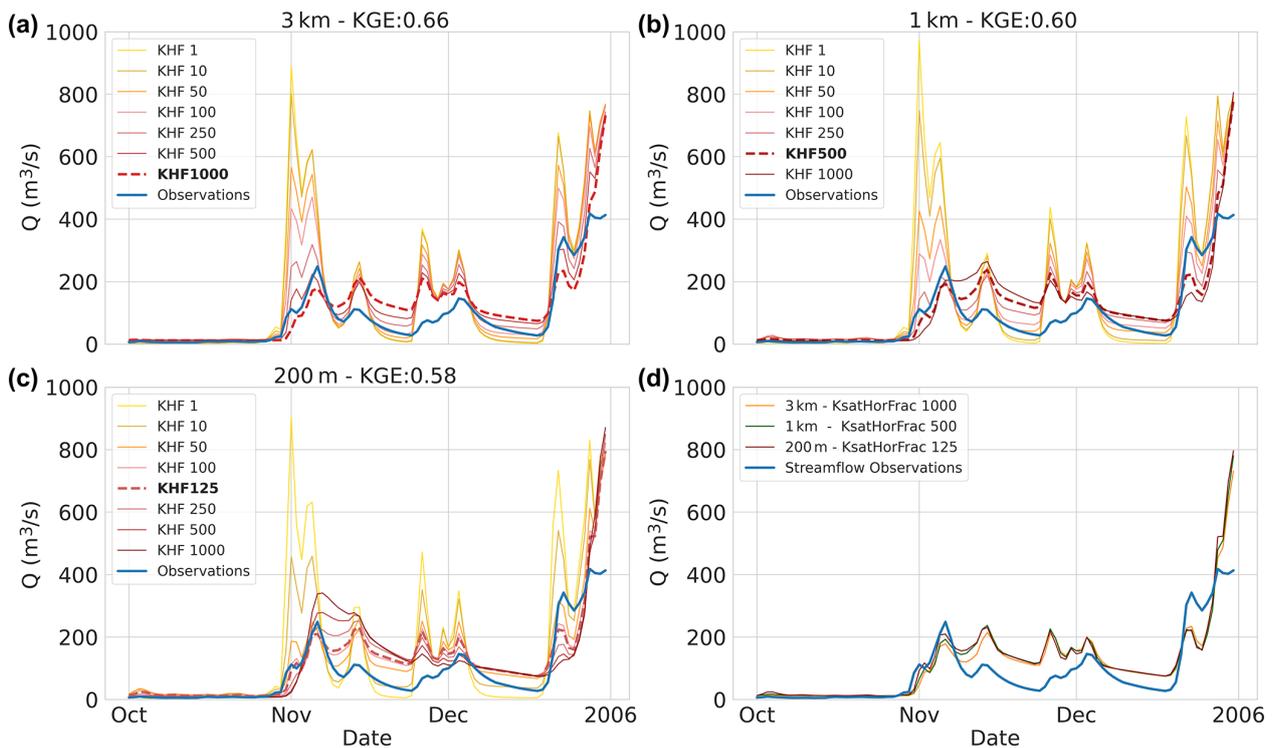


Figure 3. Calibration period results – the calibration interval of the KsatHorFrac (KHF) parameter for the three model instances at the basin outlet (ID:14301000) of the final calibration-period year: (a) 3 km model instance, (b) 1 km model instance, and (c) 200 m model instance. Values range from 1 to 1000 (yellow to red) and only a sub-selection is shown. Best-performing calibration values are indicated with a dotted red line and streamflow observations in blue. (d) Best streamflow estimates of the 3 km (orange), 1 km (green), and 200 m (red) model instances.

olution instance (3 km) in many cases. Of the 454 basins, the 3 km instance has the lowest peak-flow estimates in 279 occurrences of which the model performs best 148 times. The example in Fig. 7b shows that best performance of the 1 km model instance often occurs in conjunction with a relatively similar performance of the 200 m model instance. In 78.7 % of the cases in which the 1 km instance is best performing, the difference in KGE score with the 200 m instance is smaller than 0.1. The final example in Fig. 7c illustrates when the finest spatial model resolution (200 m) is best performing and the coarsest (3 km) least performing. The 200 m model instance best captures peak flow and the receding limbs of the hydrograph.

3.2.3 Streamflow estimates of model instances

The KGE score results for the evaluation period are shown in Fig. 8a. The KGE score distribution of the mean of 36 hydrological models from Knoben et al. (2020) is included and referred to as “MARRMoT mean results”. It is worth noting that the comparison between studies is not one-on-one due to differences in model run periods, forcing, and numerical solvers. We can, however, obtain information about general model performance between both studies.

The mean KGE score distribution of the MARRMoT models (Fig. 8, blue) of Knoben et al. (2020) is close to the mean of the distributions of the three model instances. Differences between study results are mainly present in the tails of the distributions. Below 0.17 of the CDF (worst 17 % of the results), the MARRMoT mean results of KGE score distribution is higher than the 1 km model instance. The MARRMoT mean results for the lower 5 % of the CDF performs better than the distributions of the three model instances. Here, the range of KGE scores is smaller for the MARRMoT mean (−1.55 to 0.09) than for the three model instances (−13.56 to 0.00). Above 0.17 of the CDF (83 % of the results), the distributions of the three model instances are higher in KGE scores than those of the MARRMoT mean.

When we only consider the `wflow_sbm` instances, approximately 64 % of the results of the model instances are higher than 0.50 KGE score, and of those, 18 % are higher than 0.75 KGE score. The distributions cross at multiple points, for example at the bottom 10 % of the distribution the 3 km instance has the highest and the 1 km the lowest KGE score. At 40 % of the distribution and lower, the 200 m instance is followed by the 1 km and 3 km instances in terms of highest KGE score. The Pearson r component of the KGE score CDF in Fig. 8b and the gamma variability ratio component in

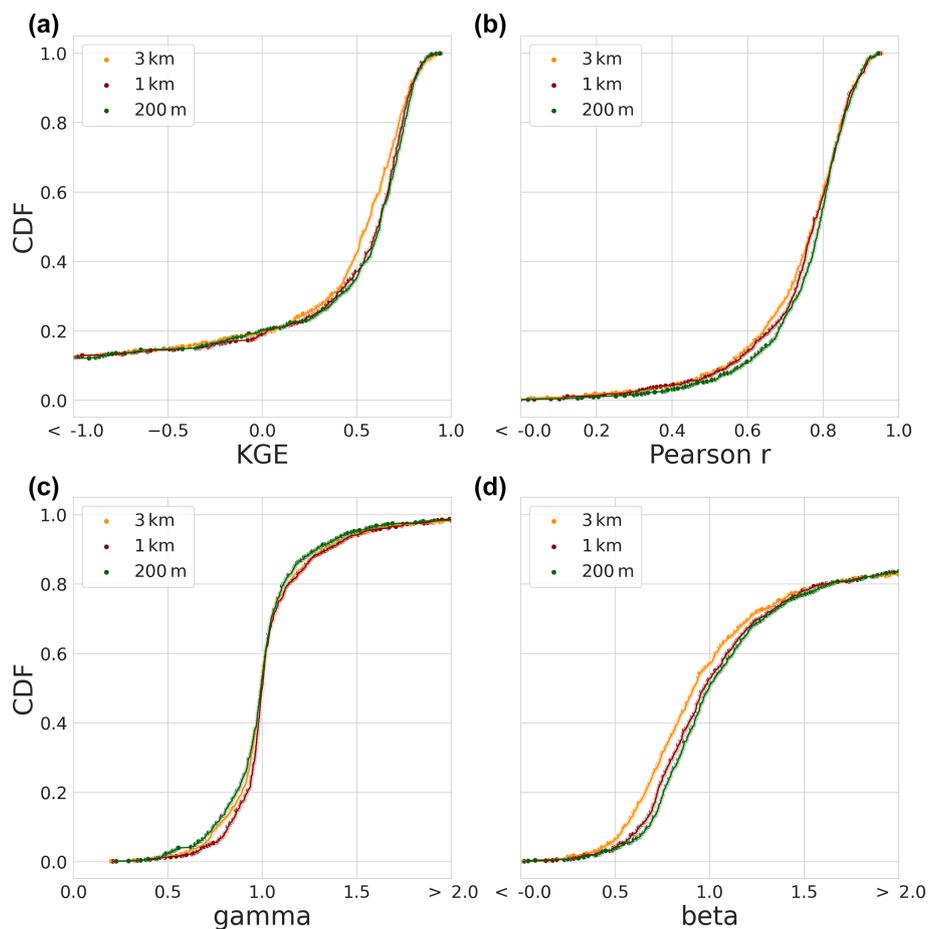


Figure 4. Calibration period results – the modified KGE score CDFs of the best-performing model runs during the calibration period and its individual components of the three model instances, i.e. 3 km (orange), 1 km (red), and 200 m (green) model instances. **(a)** The CDF of the modified KGE score. **(b)** The CDF of the Pearson r correlation component. **(c)** The CDF of the gamma (variability ratio) component. **(d)** The CDF of the beta (bias ratio) component.

Fig. 8c show small differences between model instances. The beta bias component in Fig. 8d shows the largest differences between model instances, especially between the 3 km, 1 km, and 200 m model instances. The bias component is the main factor for the differences in the overall KGE score.

Next, we apply the KS statistic to test whether the CDF of the model instances statistically differ from each other, for a given p -value of 0.05. The KS-test results in Table 2 show that the difference between 3 km and 200 m model instances is statistically relevant at a p -value of 0.02. On a large sample, this means that increasing the spatial model resolution from 3 km to 1 km or 1 km to 200 m does not lead to significant differences in streamflow performance. When changing resolution from 3 km to 200 m, the distribution of KGE scores is significantly different ($p < 0.05$) according to the KS-test.

Table 2. The Kolmogorov–Smirnov statistic results and the corresponding p -value. The results are based on the difference between the KGE distributions of the three model instances: 3 km, 1 km, and 200m.

CDFs	Kolmogorov–Smirnov Statistic	p -value
3 km–1 km	0.08	0.14
1 km–200 m	0.05	0.82
3 km–200 m	0.11	0.02

3.2.4 Objective function uncertainty

In addition to the streamflow evaluation, we conducted a sampling uncertainty assessment of the KGE objective function using bootstrap and jackknife-after-bootstrap methods. The results of this assessment for each of the model instances are shown in Fig. 9. The results for the three model instances are very similar. The tolerance-interval results, denoted with

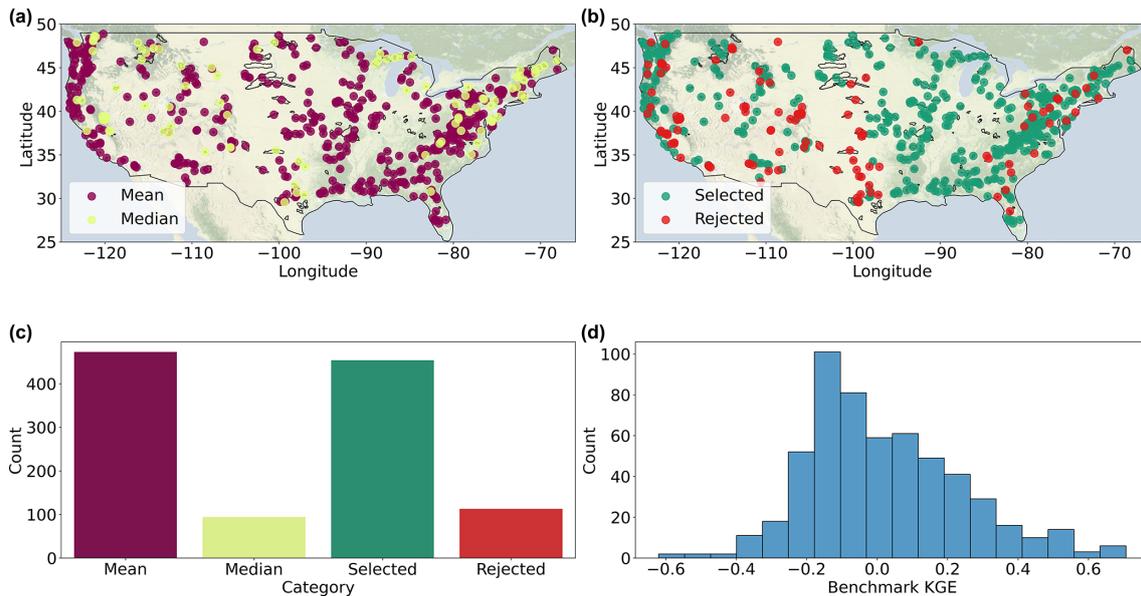


Figure 5. Evaluation period results. (a) The best-performing type of 10-year calendar day climatology, either mean (purple) or median (yellow). (b) The spatial distribution of accepted (green) and rejected (red) basins based on the benchmark. (c) Overview of the amount of basins that are accepted or rejected and the best-performing benchmark type. (d) The 2012 KGE score distribution of the best-performing benchmark type. Basemaps are made with Natural Earth.

Table 3. Sample uncertainty analysis results per quarter of the total percentage of the modified KGE cumulative distribution function (CDF) of the evaluation period. The mean of the three model instances' results is calculated based on the tolerance interval, jackknife standard error, and the bootstrap standard error, for each quarter of the total percentage.

CDF	Mean tolerance (p95–p05)	Mean 2× std err jackknife	Mean 2× std err bootstrap
0.00–0.25	6.35	5.11	4.00
0.26–0.50	0.29	0.18	0.21
0.51–0.75	0.19	0.12	0.13
0.76–1.00	0.14	0.09	0.10

the black lines, show that approximately 100 basins have a KGE sampling uncertainty of 0.1 or lower, and approximately half of the basins have values of 0.2 or lower. Half of the basins show high KGE uncertainty of more than 0.2 with approximately 80 basins surpassing 0.5 KGE for all model instances.

We project the sample uncertainty results on the CDF of the evaluation period (Fig. 8) by calculating the mean of the tolerance interval, jackknife standard error, and bootstrap standard error for each of the three model instances per quarter of the total CDF. The results in Table 3 show slightly higher results for quarters of the CDF for the tolerance interval relative to the jackknife and bootstrap standard errors. The lower tail of the CDF contains the highest average values for the three sample uncertainty statistics, and the upper

tail contains the lowest average values, indicating that sample uncertainty is high at low KGE values and vice versa.

3.2.5 Spatial distribution of evaluation period results

The CDF does not provide information at a basin level. To gain insight into the spatial distribution of the KGE scores of the model instances, Fig. 10 shows the KGEs of the streamflow estimates plotted on a map of the CONUS domain. The minimum KGE scores of 0.50 to 0.89, shown in Fig. 10a, are found in the Pacific Northwest, South Atlantic, Appalachia, and northeast of CONUS. The KGE scores lower than -0.41 are found throughout CONUS. The highest KGE scores in Fig. 10b are located in the Northwest, Rocky Mountains, and Appalachia. These regions are characterized as steep-sloping headwater basins. Figure 10c shows that there are large local streamflow discrepancies of more than 1.00 KGE score. These are mainly found in the Pacific Southwest, the Southwest, and the Midwest. These regions span a wide range of hydro-climatic diverse basins. The average KGE score difference is 0.22. Figure 10d shows the best-performing model instance for each of the 454 selected basins. Although regions show clusters in best-performing model instances, there is no overall geographical trend in results. Best performances for the 1 km model instance are generally close to basins where the 200 m model instance performs best. The 3 km model instance shows some clusters in the Southwest and Pacific Northwest. The Rocky Mountains contains the best-performing model instances at 200 m and the Appalachia

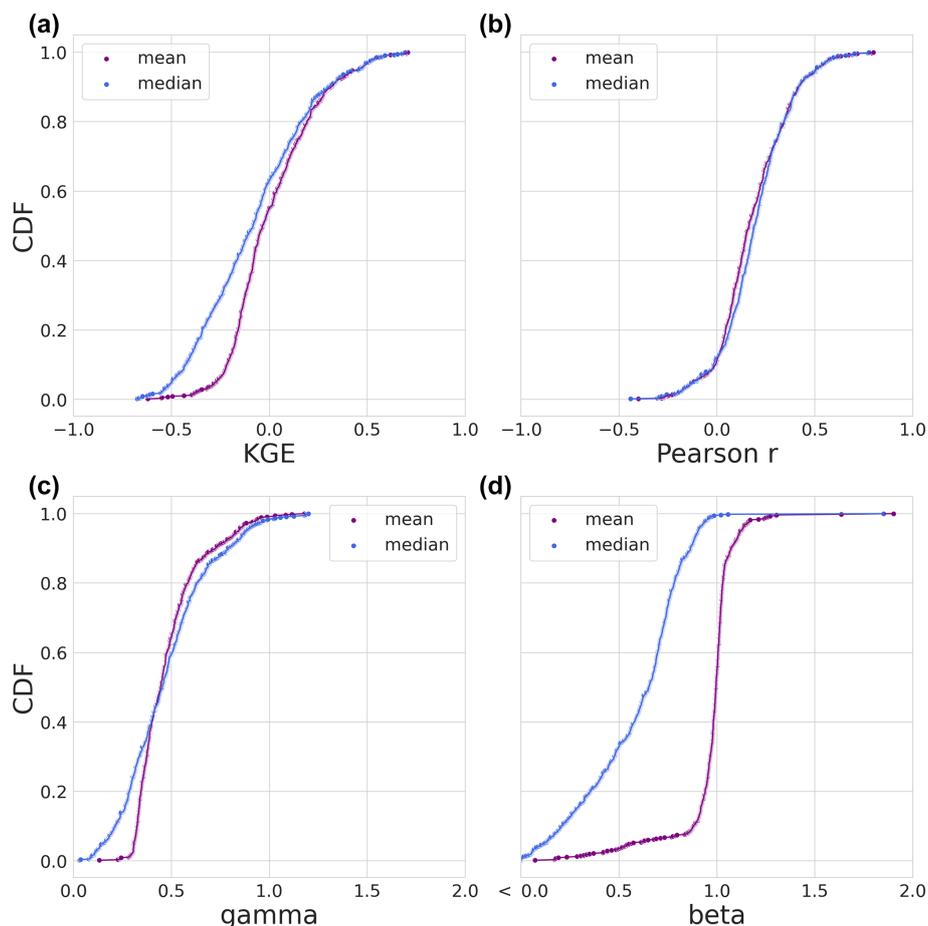


Figure 6. Evaluation period results. The CDFs of the modified KGE score and its three individual components of the statistical benchmark during the evaluation period. The mean is shown in purple and the median statistical benchmark in blue. **(a)** The CDF of the modified KGE score. **(b)** The CDF of the Pearson r correlation KGE component. **(c)** The CDF of the gamma (variability ratio) KGE component. **(d)** The CDF of the beta (bias ratio) KGE component.

a shows a mixture between the 1 km and 200 m model instances.

3.3 The effect of spatial scale on terrain characteristics

We illustrate the effect of spatial scaling on the parameter set of the three model instances by showing the difference in topography and drainage density for three basins. To avoid presentation bias, the basins were sampled based on poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). Figure 11a shows the PDF of the height distributions of the model instances for each of the three basins, Fig. 11b shows the slope distribution, and Fig. 11c the profile curvature distributions.

The height distribution of the DEM in Fig. 11a shows, most clearly for basin ID 02231342, how the representation of the highest altitudes is underestimated by the 3 km model instance (orange) compared to the 200 m instance (green), and to a lesser extent the 1 km instance (red). Essentially, at

coarser spatial resolution the terrain is flattened at high altitudes. An opposite effect is shown in this basin for the lower altitudes where the finer-resolution instances better capture gentle slopes that are flattened at coarse resolution. This effect is also detectable in the slope and profile curvature PDFs shown in Fig. 11b and c. As can be expected from the height distributions, the slope of the 200 m instance has more gentle and steep sloping topography than the 1 and 3 km instances. This is shown by the narrower slope distribution for the coarse spatial resolution that broadens with finer resolution. The differences in the mean slope of the basins between model instances is marginal, e.g. 0.00019 m m^{-1} for basin ID 02231342. The profile curvature in Fig. 11c indicates whether a slope is linear (values close to 0), concave (values smaller than 0), and convex (values larger than 0). The 3 and 1 km instances show similar slope geometries with the 3 km instance having slightly more linear slopes. At the finest resolution (200 m) the slope's geometry shifts from linear slopes to either convex or concave curvature profiles.

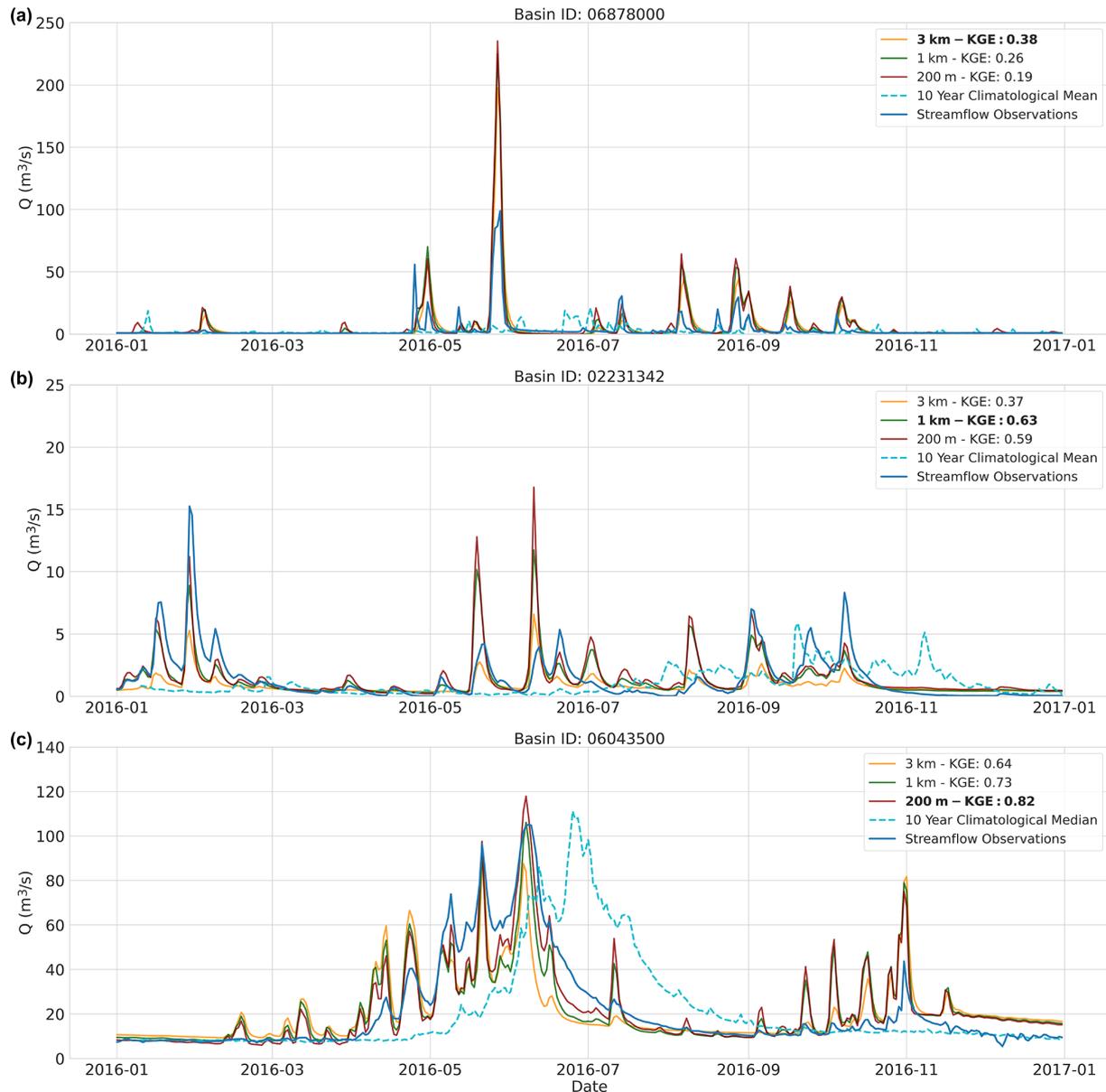


Figure 7. Evaluation period results. Three example hydrographs showing the last year of the evaluation period. The 3 km (orange), 1 km (red), and 200 m (green) model instance streamflow estimates at the basin outlet are shown. The streamflow observations are shown in blue and the 10-year calendar-day climatology of the statistical benchmark is shown in dotted cyan. **(a)** Basin ID: 06878000. **(b)** Basin ID: 02231342. **(c)** Basin ID: 06043500.

In addition to topography, we calculated the drainage density for each of the model instances defined as total river length divided by basin area. The results in Table 4 show small differences between the model instances for each of the three basins.

4 Discussion

4.1 Benchmarks

We applied an initial statistical benchmark, based on streamflow observations for basin selection, to identify the basins in which streamflow estimates are deemed adequate for further analysis. This does not imply that excluded basins are less relevant. Instead, it implies that an in-depth model assessment is required to understand why the model is not able to simulate adequate streamflow estimates in these basins. The

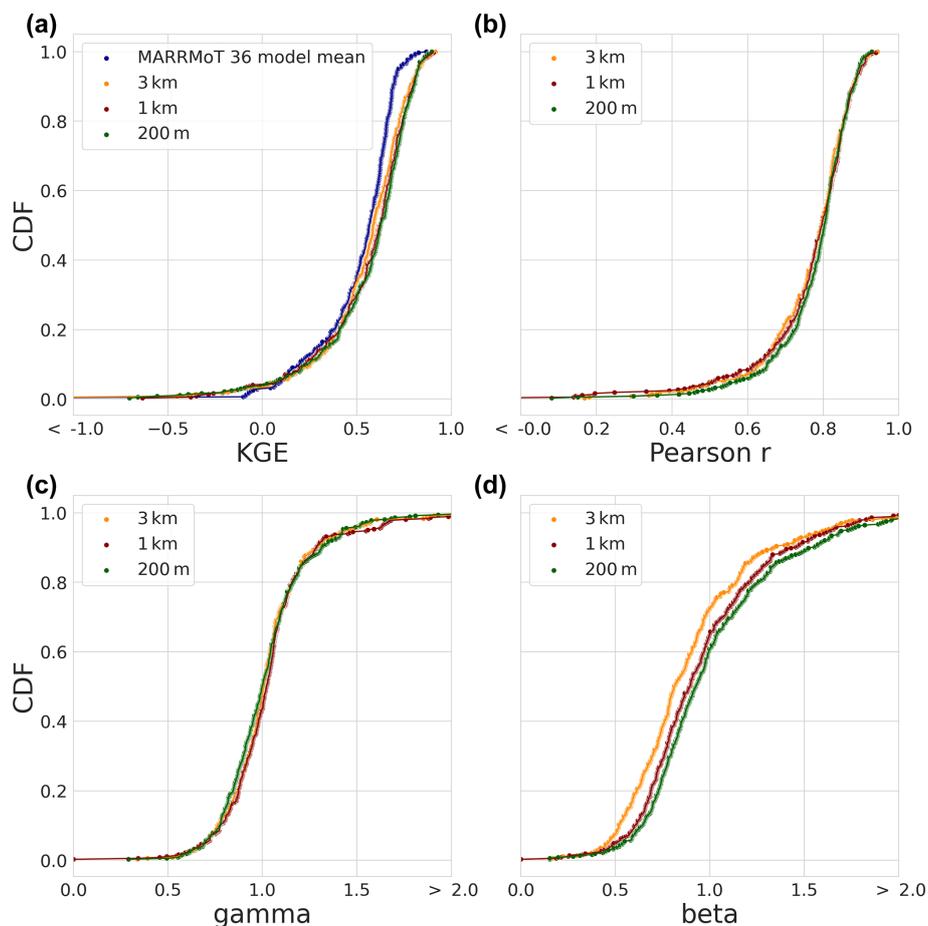


Figure 8. Evaluation period results. The CDF based on the modified KGE scores and its three individual components for the 454 selected basins. (a) The modified KGE score for the MARRMoT 36 mean results (blue), the 3 km (yellow), 1 km (red), and 200 m (green) model instances. (b) The Pearson r correlation component. (c) The gamma variability component. (d) The beta bias component.

Table 4. The drainage density defined as total river length divided by basin area for the three example basins.

Basin ID	3 km drainage density [m m ⁻²]	1 km drainage density [m m ⁻²]	200 m drainage density [m m ⁻²]
06878000	0.0133	0.0141	0.0143
02231342	0.0056	0.0059	0.0063
06043500	0.0112	0.0123	0.0126

CDFs of the KGE score and its components based on the statistical benchmarks in Fig. 6 show that the benchmark is relatively easy to beat given the KGE score distributions. We find, as one might expect, that the beta bias component has values close to 1 for the mean benchmark while this is not the case for the median benchmark. Bias values that are not close to 1 for the mean benchmark indicate that the flow regime changes from year to year. In most cases the hydrological

model simulations were able to capture this change better than the benchmark. This shows that the model is able to capture yearly variability and is not overfitting due to extensive calibration. In addition to the benchmark, we added a layer of context by including results from the study by Knoben et al. (2020). This is an imperfect comparison due to differences in model inputs, numerical solvers, and simulation period. However, the results do provide information on general model performance. The results show that the wflow_sbm streamflow estimates are inline with estimates of the mean from the 36 MARRMoT models. The spread of results is smaller for the 36 MARRMoT models which is due to averaging and likely the more extensive calibration routine of the conceptual models. It also implies that when only streamflow at the basin outlet is under consideration, users should carefully consider various model structures before model selection due to the small differences in results at the system scale.

Other studies have conducted large domain-modelling efforts with CONUS as case study area (e.g. Mizukami et al.,

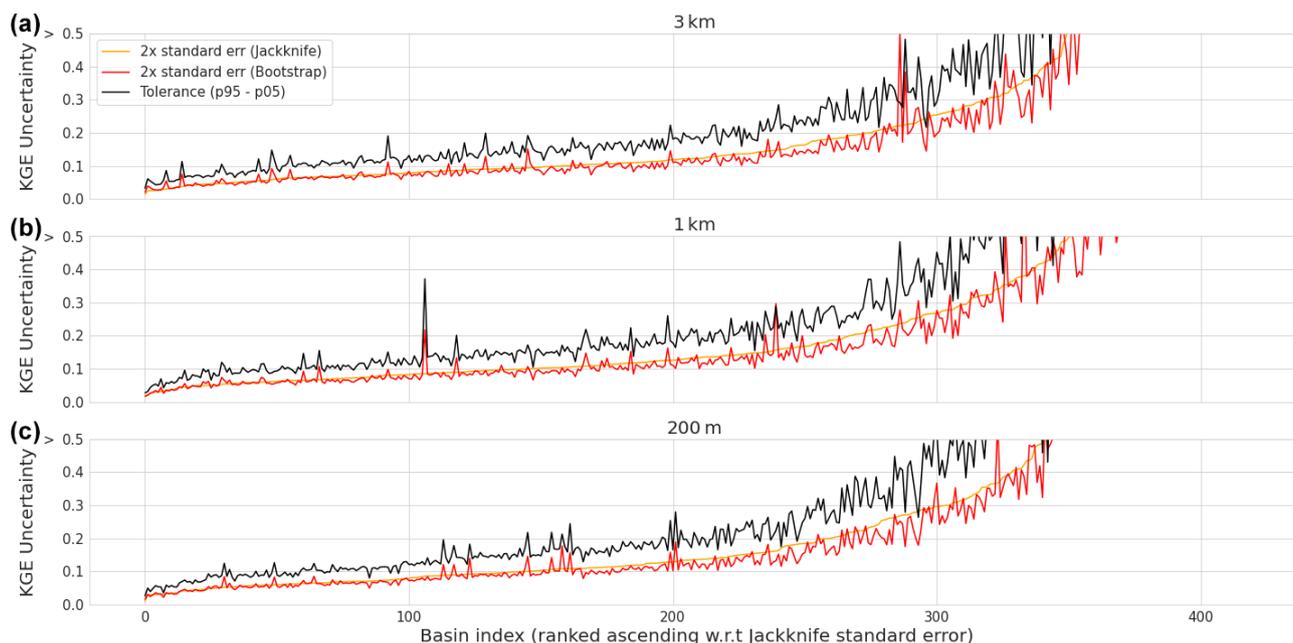


Figure 9. Evaluation period results. The bootstrap and jackknife-after-bootstrap results of the sampling uncertainty of the KGE score. The $2\times$ standard error of the jackknife method is shown in orange, the $2\times$ standard error of the bootstrap method is shown in red, and the tolerance interval obtained by subtracting the 5th percentile from the 95th percentile is shown in black. The horizontal axis shows the basin index ranked w.r.t. the jackknife standard error and the vertical axis shows the modified KGE score sampling uncertainty. (a) The 3 km model instance results. (b) The 1 km model instance results. (c) The 200 m model instance results.

2017; Rakovec et al., 2019). However, these are hard to compare with the results from this study as they did not use the same basins. To improve future comparative work, we advocate for the creation of model output storage guidelines that use the CAMELS data set as case study area. These guidelines should encompass the differences between hydrological models, such as distributed and non-distributed modelling grids. A first step is the inclusion of distributed data sources in the CAMELS data set, e.g. meteorological data. This can be extended by including model evaluation products such as snow cover and soil moisture for further benchmarking. We further propose the use of a model experiment environment, such as eWaterCycle (Hut et al., 2021) to generate model results. This allows for similar pre-processing of inputs, standardization of outputs, and reproducible modelling studies. An example of how to apply these steps using the eWaterCycle platform is provided in the Jupyter Notebooks that supplement this publication (<https://doi.org/10.5281/zenodo.5724512>). The ease of setting up a model experiment and storing output is an incentive for users to store model results while conducting extensive modelling studies, even when results are not deemed not suitable for publication yet but might still benefit the community.

4.2 Streamflow estimates and uncertainty

At the start of the study we hypothesized that differences between model instances would be small due to quasi-scale invariant parameter sets and constant hydrological process descriptions within the model. The results of the calibration period in Fig. 4 and the evaluation period in Fig. 8 show that this is the case for the KGE score distributions based on streamflow estimates at the basin outlet. Although the differences are small, the crossing of the distribution lines is a strong indication that there is disagreement about KGE scores between model instances and that there is no single instance outperforming another consistently. The largest difference between distribution for both periods is found in the beta bias component of the KGE score. The benchmark selection mainly excluded the basins that contained large bias values during the calibration period. For the bias component, the 3 km instance deviates from the 1 km and 200 m instances. This shows that the total volume of streamflow differs for the 3 km instance. This difference was confirmed by testing the statistical significance of the differences between distributions of the model instances by applying the KS-statistic on the evaluation period results. Here, a statistical difference between distributions is only found for the model-instance combination of 3 km and 200 m Table 2.

This study applies a single objective function, the modified KGE (Kling et al., 2012), to determine simulated streamflow adequacy for the model calibration and evaluation time pe-

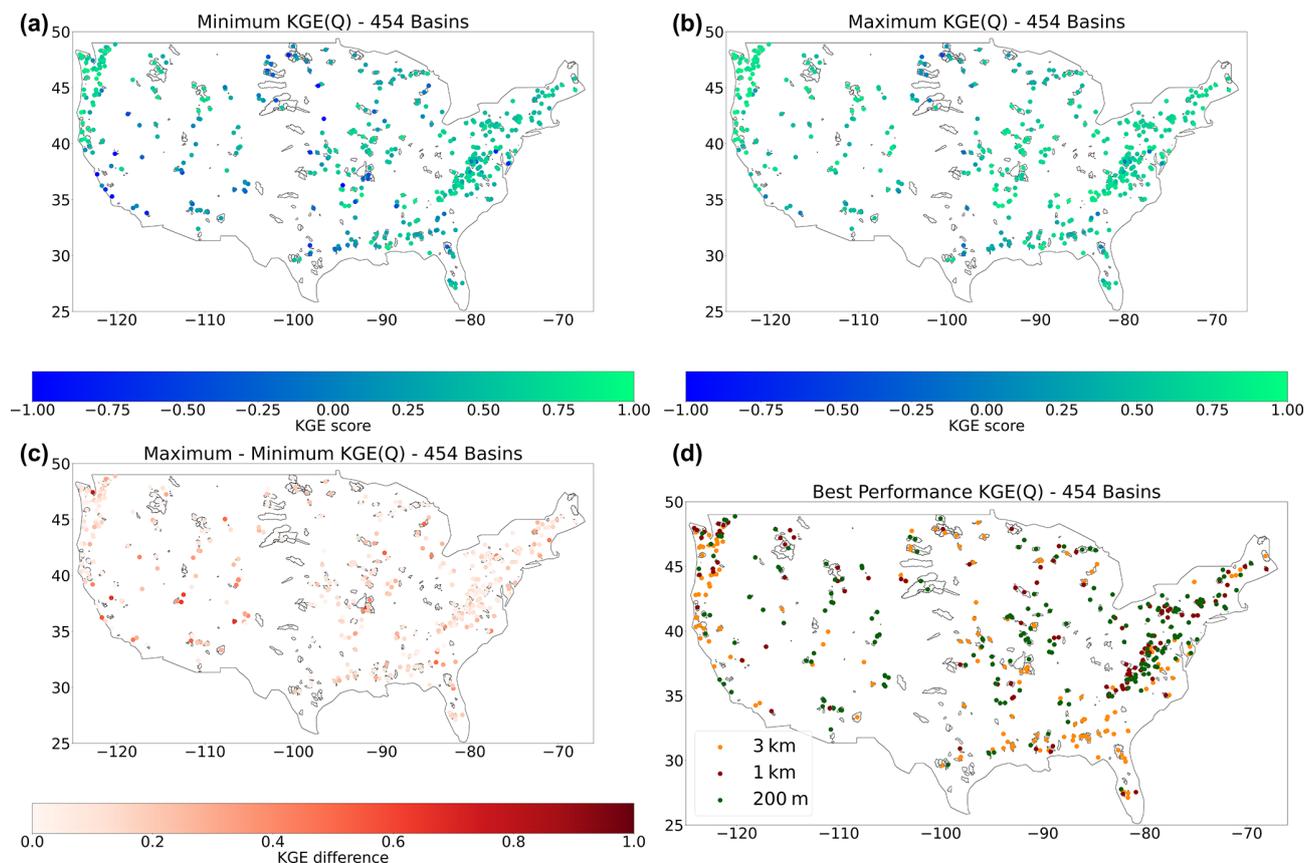


Figure 10. Evaluation period results. (a) Minimum KGE score of the model instances. (b) Maximum KGE score of the model instances. (c) The difference between minimum and maximum KGE scores. (d) Best performing model instance based on KGE score for the evaluation period with 3 km in orange, 1 km in red, and 200 m in green.

riods. To provide the reader with more context we have included the 2009 KGE (Gupta et al., 2009), modified KGE, non-parametric KGE (Pool et al., 2018), and NSE objective function results in the data repository and in Fig. A2. The hypothesis testing based on conclusions were not affected by the type of KGE objective function. We selected the modified KGE objective function as it is less influenced by extreme combinations of simulated streamflow, observed streamflow, and structural error in the meteorological input. The single objective function for a whole period approach is limited and can be improved by first determining the objective function for individual years and then averaging it for the whole period (Fowler et al., 2018). In addition, as stated in Clark et al. (2021), it is important to determine the sampling uncertainty of objective functions to avoid incorrect conclusions at the system scale. Following their methodology, we investigated the sampling uncertainty by applying bootstrap and jackknife-after-bootstrap methods (Fig. 9). Approximately half of the basins have KGE uncertainties of 0.2 or lower based on the tolerance interval, and half of higher than 0.2 with values surpassing 0.5 KGE uncertainty. The lower 25 % of the CDF (Fig. 8 contains the highest sample uncertainty

and the upper 25 % the lowest (Table 3). Given these results, we argue that the differences found using the KS-statistic are easily within the sampling uncertainty margins and therefore not valid to base conclusions on, even more so when considering the model uncertainty and the observation uncertainty. This demonstrates the drawbacks of conducting large-sample assessments and the sensitivity to sampling uncertainty of the results.

4.3 Relative model instance differences

We recognize that large-sample assessments obscure variations in simulations between instances due to the sample size. On a basin level we find that local variations due to spatial resolution are in effect throughout the domain. This is depicted by the differences between the KGE scores of the model instances (Fig. 10c). On average, there is a 0.22 KGE score difference with extremes of more than a 0.5 KGE score difference at multiple basins. In this case, we are interested in the relative differences between the instances to understand the effect of varying spatial resolution. According to Oreskes et al. (1994), this is the only form of validation that is actually possible. Therefore, we argue that the differences are

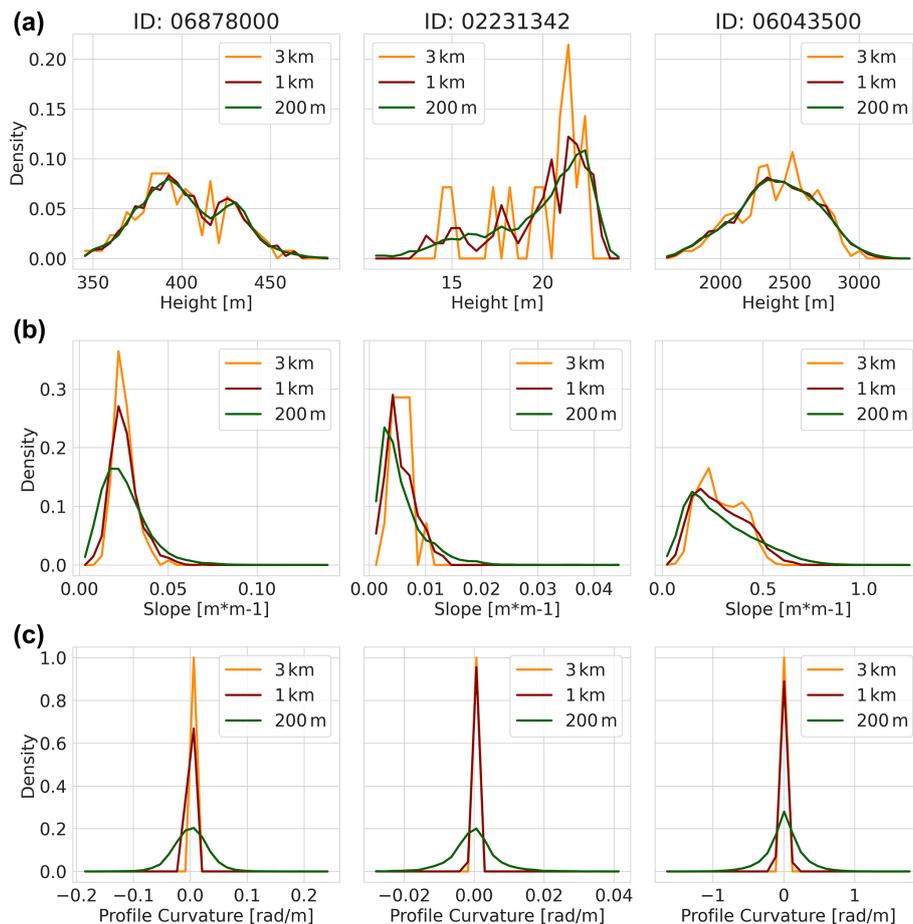


Figure 11. Three example basins that represent poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). **(a)** The PDF of height distribution for the 3 km (orange), 1 km (red) and 200 m (green) model instances. **(b)** The PDF of the slope distributions of the model instances per basin. **(c)** The PDF of profile curvature distributions of the model instances per basin. Values that equal 0 indicate linear-slope geometry, smaller than 0 concave-slope geometry, and larger than 0 convex-slope geometry.

large locally, even though they might be within the sampling uncertainty range.

We find that the 1 km instance performs best in basins where the difference between minimum and maximum KGE score of the three instances is small (Fig. 10c). We partly attribute this to the calibration routine finding a better optimal parameter value (KsatHorFrac) since results are often close to those of the 200 m instance. For the best-performing 3 km or 200 m model instances (Fig. 10d), there are only small geographical trends of best model performance in the South and Appalachia. This information is valuable for future research that will conduct a more in-depth assessment of the internal states and fluxes because now we know where the effects of varying grid-cell size is small or large and which model performs best.

We conducted a terrain analysis (Fig. 11) to identify changes in terrain characteristics due to spatial resolution that might explain the differences between streamflow estimates of model instances. Minor effects are depicted by the profile

curvatures present at 200 m resolution. Slopes are less linear at fine resolution than at coarse resolution. The effect on the hydrological response is, however, expected to be small as stated by Bogaart and Troch (2006). Similarly, small changes are found for the differences in drainage density between model instances (Table 4). This confirms that the drainage network upscaling method of Eilander et al. (2021) is (almost) consistent across spatial scales.

Larger differences between model instances are found for the height distribution of the DEM, which is flattened at coarse resolution compared to finer resolution. This introduces changes in snow dynamics at high altitudes due to the use of the temperature degree-day method by the hydrological model. The resulting effect on streamflow estimates depends on the relative contribution of snowmelt. Although marginal at a basin level, the difference in slope between instances is expected to effect the partitioning of the lateral fluxes of the wflow_sbm model since the lateral connectivity between grid cells is slope-driven. An increase in slope

would lead to larger lateral fluxes and vice versa. Increasing spatial resolution, aggregating the DEM, results in a broader distribution of slopes that affects the volume and timing of streamflow estimates. The effect of terrain smoothing has been reported by Shrestha et al. (2015) and they found this to increase overland flow lateral flux. An in-depth assessment of the internal states and fluxes of the model instances is required to determine whether these components are the main cause for the differences in streamflow estimates.

We applied the same meteorological forcing products and pre-processing routine for each model instance. This ensured that the total volume of precipitation remained consistent across scales. A coarse grid cell contains a volume of precipitation that is equally redistributed over the equivalent amount in size of finer grid cells. In reality, this redistribution of water might not be equal across the finer grid cells, and thus scaling behaviour is introduced due to the locality of precipitation. This has an effect on the streamflow estimates as the locality of precipitation directly influences hydrological processes that are dominant at different locations (e.g. hill slope). Additionally, due to the large difference between native data and model instance resolution, it is likely that the effects of disaggregation of precipitation and temperature lapsing are main drivers for differences in streamflow estimates between model instances. However, Shuai et al. (2022) found that by comparing the same forcing product with native implementations at various spatial resolutions the effect on streamflow estimates was relatively small. This was not the case for distributed variables in the basin (e.g. snow water equivalent). It is of interest to investigate this effect on streamflow estimates and to determine the role of native spatial meteorological forcing resolution.

4.3.1 Computational cost

When we consider the increase in streamflow-based model performance as opposed to computational cost, we find that it does not scale linearly with the amount of grid cells in the basin due to lateral connections in the hydrological model. The average non-parallel run time of the 3 km instance is 157 s while that of the 200 m instance 12 120 s with an average grid-cell number difference of 28 872 cells. These results point toward the importance of conducting an initial spatial model resolution assessment at the start of large-sample assessments as it avoids sub-par or computationally expensive model runs. Note that this kind of information can stimulate scientific and/or computational developments, e.g. in the meantime the wflow code was rewritten in Julia (van Verseveld et al., 2021), roughly increasing performance by a factor of 3 while other improvements (threading, mpi) are being implemented. There are alternative approaches to the spatial discretization of basins that are computationally very efficient, such as the vector-based configurations that have the added benefit to better capture topographic details and

are less influenced by native forcing resolution, for example (Gharari et al., 2020).

4.4 Outlook

The results from this study help model developers with model refinement by providing them with an understanding of where and under which circumstances differences due to spatial scaling occur. Based on the aggregated domain and basin level results we can conclude that increasing spatial resolution does not necessarily lead to better streamflow estimates at the basin outlet. The implications of the results for the user are that caution is advised when interpreting high-resolution model outputs as this does not directly translate into better model performance. Moreover, the computational cost of increasing model resolution is not always warranted compared to increase in streamflow estimate-based model performance.

We conducted this study as an initial assessment to be followed up with studying scaling effects in distributed hydrological models. As the sampling uncertainty results showed, it is very hard to draw conclusions from a large sample and future research should therefore consider a smaller subset of basins to explore scaling effects in more detail. In this study, we did not investigate individual basins to avoid biased selection of case study areas. We suggest that future work investigates the basins that show large or small differences in model performance, lateral fluxes, and effects of terrain aggregation to be part of this subset. In addition, the evaluation should go beyond streamflow by using multiple evaluation data products (e.g. soil moisture, evaporation, gravitational anomaly, see (Guse et al., 2021) for a recent overview). Conclusions might then be drawn on whether increasing spatial resolution leads to increased model fidelity or not. This should be tested using multiple forcing data sets to evaluate the robustness of the conclusion. With the inclusion of multiple timescales as discussed in Melsen et al. (2016), more information can be obtained about the linearity of hydrological process descriptions in the model.

5 Conclusions

The aim of this study was to analyse the effects that varying spatial resolution has on the streamflow estimates of the distributed wflow_sbm hydrological model. Distributions of model instance KGE score results were tested for significant differences as well as the sampling uncertainty. A spatial distribution assessment was conducted to derive spatial trends from the results. The main findings of the study are the following:

- The difference in the distributions of streamflow estimates of the wflow_sbm model derived at multiple spatial grid resolutions (3 km, 1 km, 200 m) is only statistically significantly different between the 3 km and

200 m model instances ($p < 0.05$). This confirms the hypothesis at an aggregated level that differences between model instances are small due to quasi-scale invariant parameter set and process descriptions that remain constant across scales in the hydrological model. However, the sampling uncertainty of the KGE score proved to be large throughout the domain. Therefore, we conclude that the minute differences found between model instances are too small to base conclusions on.

- Results show large differences in maximum and minimum KGE scores with an average of 0.22 between model instances throughout CONUS. This provides valuable information for follow-up research based on the locality of relative model scaling effects.
- There is no single best-performing model resolution across the domain. Finer spatial resolution does not always lead to better streamflow estimates at the outlet.
- Changes in terrain characteristics due to varying spatial resolution influence the lateral flux partitioning of the wflow_sbm model and might be an important cause for differences in streamflow estimates between model instances.

This study indicated where locality in results are strong due to varying spatial resolution. Future research should conduct an in-depth assessment of basins where differences in streamflow estimates and lateral fluxes are large due to spatial scale. This will lead to a better understanding of why and under which conditions locality in spatial scaling-related issues occur.

Appendix A

A1 ERA5 and MSWEP precipitation forcing comparison

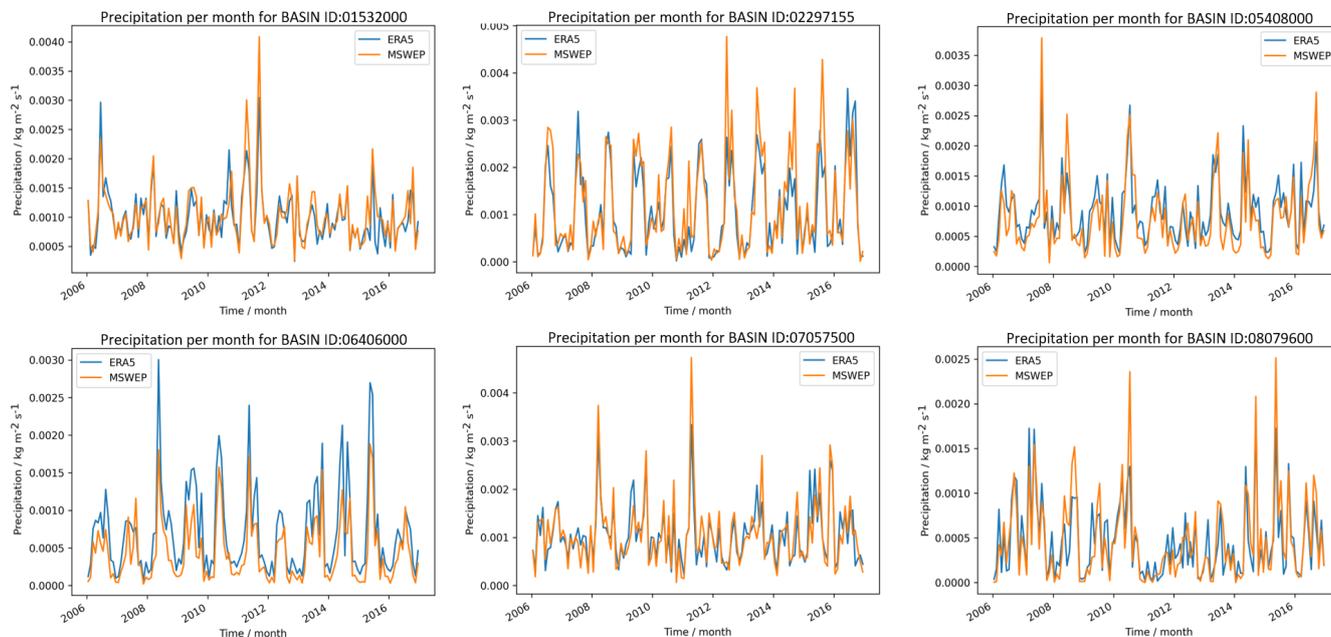


Figure A1. Evaluation period results. ERA5 and MSWEP forcing comparison for six basins in the CAMELS data set. Monthly precipitation values for the evaluation period are shown for ERA5 in blue ERA5 and for MSWEP in orange.

Table A1. Evaluation period results. Comparison of evaluation period objective function results of the 3 km wflow_sbm instance based on the ERA5 and MSWEP forcing data sets.

BASIN ID	Resolution	MSWEP-KGE 2012	ERA5-KGE 2012	MSWEP-NSE	ERA5-NSE
01532000	3 km	0.20	0.19	0.67	0.65
02297155	3 km	0.23	0.04	0.34	0.06
05408000	3 km	0.35	-0.03	0.60	0.39
06406000	3 km	-1.69	-6.19	0.21	0.15
07057500	3 km	0.68	0.53	0.56	0.35
08079600	3 km	-5.65	-8.13	-0.23	-0.49

A2 CDFs of multiple objective functions

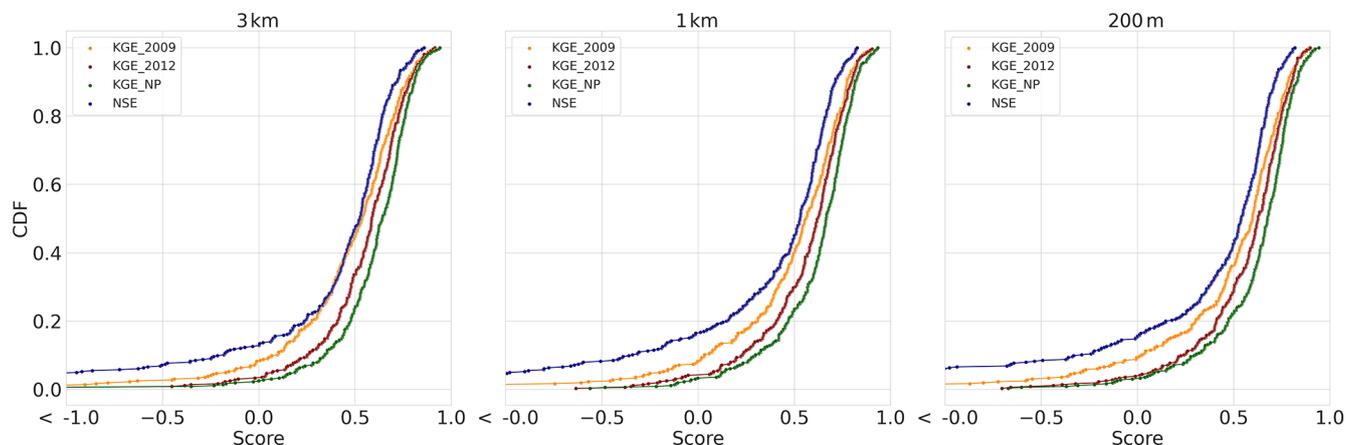


Figure A2. Evaluation period results. CDFs of multiple objective functions for the three model instances with KGE 2009 in orange, KGE 2012 in red, KGE NP in green, and NSE in blue.

Code and data availability. The software that supplements this study is available at https://github.com/jeromaerts/eWaterCycle_example_notebooks (last access: 5 June 2022.) or <https://doi.org/10.5281/zenodo.5724512> (Aerts et al., 2021a). The data that supplement this publication are available at <https://doi.org/10.5281/zenodo.5724576> (Aerts et al., 2021b).

Author contributions. JPMA wrote the publication. JPMA, WJvV, AHW, and PH did the conceptualization of the study. JPMA, ND, and PH developed the methodology. JPMA, WJvV, AHW, and PH conducted the analyses. RWH, NCvdG, WJvV, AHW, and PH did an internal review. RWH, ND, and NCvdG are PIs of the eWaterCycle project.

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

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We would like to thank the anonymous reviewer and Shervan Gharari for their valuable feedback that helped to improve this manuscript. 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.

Financial support. This research has been supported by the Netherlands eScience Center (grant no. 027.017.F0).

Review statement. This paper was edited by Efrat Morin and reviewed by Shervan Gharari and one anonymous referee.

References

- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrol. Earth Syst. Sci.*, 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- Aerts, J. P. M.: eWaterCycle_example_notebooks (Version 1), Zenodo [code], <https://doi.org/10.5281/zenodo.5724512>, 2021a.
- Aerts, J. P. M.: Wflow SBM streamflow estimates for CAMELS data set (Version 1), Zenodo [data set], <https://doi.org/10.5281/zenodo.5724576>, 2021b.
- Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, *Bull. Am. Meteorol. Soc.*, 100, 473–500, <https://doi.org/10.1175/BAMS-D-17-0138.1>, 2019.
- Bell, V. A., Kay, A. L., Jones, R. G., and Moore, R. J.: Development of a high resolution grid-based river flow model for use with regional climate model output, *Hydrol. Earth Syst. Sci.*, 11, 532–549, <https://doi.org/10.5194/hess-11-532-2007>, 2007.
- Benning, R.: Towards a new lumped parameterization at catchment scale, PHD thesis, Thesis, University of Wageningen, the Netherlands, <http://edepot.wur.nl/216531IID> (last access: 28 November 2021), 1994.
- Beven, K. J. and Cloke, H. L.: Comment on “Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring

- Earth's terrestrial water" by Eric F. Wood et al., *Water Resour. Res.*, 48, W01801, <https://doi.org/10.1029/2011WR010982>, 2012.
- Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E. F.: Hyper-resolution global hydrological modelling: what is next?, *Hydrol. Process.*, 29, 310–320, <https://doi.org/10.1002/hyp.10391>, 2015.
- Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: A review, *Hydrol. Process.*, 9, 251–290, <https://doi.org/10.1002/hyp.3360090305>, 1995.
- Bogaart, P. W. and Troch, P. A.: Curvature distribution within hillslopes and catchments and its effect on the hydrological response, *Hydrol. Earth Syst. Sci.*, 10, 925–936, <https://doi.org/10.5194/hess-10-925-2006>, 2006.
- Booij, M.: Impact of climate change on river flooding assessed with different spatial model resolutions, *J. Hydrol.*, 303, 176–198, <https://doi.org/10.1016/j.jhydrol.2004.07.013>, 2005.
- Brakensiek, D., Rawls, W., and Stephenson, G.: Modifying SCS hydrologic soil groups and curve numbers for rangeland soils, American Society of Agricultural Engineers, St. Joseph, MI, USA, ASAE Paper No. PNR-84-203, 1984.
- Bras, R. L.: Complexity and organization in hydrology: A personal view, *Water Resour. Res.*, 51, 6532–6548, <https://doi.org/10.1002/2015WR016958>, 2015.
- Brooks, R. H. and Corey, A. T.: Hydraulic Properties of Porous Media, *Hydrology Papers*, Colorado State University, Fort Collins, Colorado, p. 37, 1964.
- Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., and Smets, B.: Copernicus Global Land Cover Layers–Collection 2, *Remote Sens.*, 12, 1044, <https://doi.org/10.3390/rs12061044>, 2020.
- Ciarapica, L. and Todini, E.: TOPKAPI: a model for the representation of the rainfall-runoff process at different scales, *Hydrol. Process.*, 16, 207–229, <https://doi.org/10.1002/hyp.342>, 2002.
- Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, *Hydrol. Earth Syst. Sci.*, 21, 3427–3440, <https://doi.org/10.5194/hess-21-3427-2017>, 2017.
- 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 Resour. Res.*, 57, e2020WR029001, <https://doi.org/10.1029/2020WR029001>, 2021.
- Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R.: A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils, *Water Resour. Res.*, 20, 682–690, <https://doi.org/10.1029/WR020i006p00682>, 1984.
- de Bruin, H. A. R., Trigo, I. F., Bosveld, F. C., and Meirink, J. F.: A Thermodynamically Based Model for Actual Evapotranspiration of an Extensive Grass Field Close to FAO Reference, Suitable for Remote Sensing Application, *J. Hydrometeorol.*, 17, 1373–1382, <https://doi.org/10.1175/JHM-D-15-0006.1>, 2016.
- Efron, B.: Jackknife-After-Bootstrap Standard Errors and Influence Functions, *J. Roy. Stat. Soc. B Met.*, 54, 83–111, <https://doi.org/10.1111/j.2517-6161.1992.tb01866.x>, 1992.
- Efron, B. and Tibshirani, R.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Stat. Sci.*, 1, 54–75, <https://doi.org/10.1214/ss/1177013815>, 1986.
- Eilander, D. and Boisgontier, H.: HydroMT, GitHub [data set], <https://github.com/Deltares/hydromt>, 2021.
- 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, *Hydrol. Earth Syst. Sci.*, 25, 5287–5313, <https://doi.org/10.5194/hess-25-5287-2021>, 2021.
- Fan, Y., Clark, M., Lawrence, D. M., Swenson, S., Band, L. E., Brantley, S. L., Brooks, P. D., Dietrich, W. E., Flores, A., Grant, G., Kirchner, J. W., Mackay, D. S., McDonnell, J. J., Milly, P. C. D., Sullivan, P. L., Tague, C., Ajami, H., Chaney, N., Hartmann, A., Hazenberg, P., McNamara, J., Pelletier, J., Perket, J., Rouholahnejad-Freund, E., Wagener, T., Zeng, X., Beighley, E., Buzan, J., Huang, M., Livneh, B., Mohanty, B. P., Nijssen, B., Safeeq, M., Shen, C., van Verseveld, W., Volk, J., and Yamazaki, D.: Hillslope Hydrology in Global Change Research and Earth System Modeling, *Water Resour. Res.*, 55, 1737–1772, <https://doi.org/10.1029/2018WR023903>, 2019.
- Feddes, R. A., Kowalik, P. J., and Zaradny, H.: Water uptake by plant roots, in: *Simulation of Field Water Use and Crop Yield*, John Wiley & Sons, Inc., New York, USA, 16–30, google-Books-ID: zEJzQgAACAAJ, 1978.
- Finnerty, B. D., Smith, M. B., Seo, D.-J., Koren, V., and Moglen, G. E.: Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs, *J. Hydrol.*, 203, 21–38, [https://doi.org/10.1016/S0022-1694\(97\)00083-8](https://doi.org/10.1016/S0022-1694(97)00083-8), 1997.
- Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function, *Water Resour. Res.*, 54, 3392–3408, <https://doi.org/10.1029/2017WR022466>, 2018.
- Garrick, M., Cunnane, C., and Nash, J. E.: A criterion of efficiency for rainfall-runoff models, *J. Hydrol.*, 36, 375–381, [https://doi.org/10.1016/0022-1694\(78\)90155-5](https://doi.org/10.1016/0022-1694(78)90155-5), 1978.
- Gash, J. H. C.: An analytical model of rainfall interception by forests, *Q. J. Roy. Meteor. Soc.*, 105, 43–55, <https://doi.org/10.1002/qj.49710544304>, 1979.
- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network, *Hydrol. Earth Syst. Sci.*, 25, 2045–2062, <https://doi.org/10.5194/hess-25-2045-2021>, 2021.
- Gharari, S., Clark, M. P., Mizukami, N., Knoben, W. J. M., Wong, J. S., and Pietroniro, A.: Flexible vector-based spatial configurations in land models, *Hydrol. Earth Syst. Sci.*, 24, 5953–5971, <https://doi.org/10.5194/hess-24-5953-2020>, 2020.
- Grayson, R. B., Moore, I. D., and McMahon, T. A.: Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes, *Water Resour. Res.*, 28, 2639–2658, <https://doi.org/10.1029/92WR01258>, 1992.

- 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, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Gupta, V. K., Rodríguez-Iturbe, I., and Wood, E. F.: Scale Problems in Hydrology, *Water Science and Technology Library*, vol. 6, Springer Dordrecht, <https://doi.org/10.1007/978-94-009-4678-1>, 1986.
- Guse, B., Kiesel, J., Pfannerstill, M., and Fohrer, N.: Assessing parameter identifiability for multiple performance criteria to constrain model parameters, *Hydrol. Sci. J.*, 65, 1158–1172, <https://doi.org/10.1080/02626667.2020.1734204>, 2020.
- Guse, B., Fatichi, S., Gharari, S., and Melsen, L. A.: Advancing Process Representation in Hydrological Models: Integrating New Concepts, Knowledge, and Data, *Water Resour. Res.*, 57, e2021WR030661, <https://doi.org/10.1029/2021WR030661>, 2021.
- Haddeland, I., Lettenmaier, D. P., and Skaugen, T.: Reconciling Simulated Moisture Fluxes Resulting from Alternate Hydrologic Model Time Steps and Energy Budget Closure Assumptions, *J. Hydrometeorol.*, 7, 355–370, <https://doi.org/10.1175/JHM496.1>, 2006.
- Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangquan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, *PLOS ONE*, 12, e0169748, <https://doi.org/10.1371/journal.pone.0169748>, 2017.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Q. J. Roy. Meteor. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Horritt, M. S. and Bates, P. D.: Effects of spatial resolution on a raster based model of flood flow, *J. Hydrol.*, 253, 239–249, [https://doi.org/10.1016/S0022-1694\(01\)00490-5](https://doi.org/10.1016/S0022-1694(01)00490-5), 2001.
- Houze Jr., R. A.: Orographic effects on precipitating clouds, *Rev. Geophys.*, 50, RG1001, <https://doi.org/10.1029/2011RG000365>, 2012.
- Hrachowitz, M. and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, *Hydrol. Earth Syst. Sci.*, 21, 3953–3973, <https://doi.org/10.5194/hess-21-3953-2017>, 2017.
- Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., Aerts, J., Albers, T., Alidoost, F., Andela, B., Camphuysen, 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, *Geosci. Model Dev.*, 15, 5371–5390, <https://doi.org/10.5194/gmd-15-5371-2022>, 2022.
- Hutton, E. W. h., Piper, M. D., and Tucker, G. E.: The Basic Model Interface 2.0: A standard interface for coupling numerical models in the geosciences, *Journal of Open Source Software*, 5, 2317, <https://doi.org/10.21105/joss.02317>, 2020.
- 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 Resour. Res.*, 56, e2019WR026807, <https://doi.org/10.1029/2019WR026807>, 2020.
- Karger, D. N., Conrad, O., Böhrner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, *Scientific Data*, 4, 170122, <https://doi.org/10.1038/sdata.2017.122>, 2017.
- Karssenbergh, D. J., Schmitz, O., Salamon, P., de Jong, K., and Bierkens, M. F. P.: A software framework for construction of process-based stochastic spatio-temporal models and data assimilation, *Environ. Modell. Softw.*, 25, 489–502, <https://doi.org/10.1016/j.envsoft.2009.10.004>, 2010.
- Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, *J. Hydrol.*, 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, *Geosci. Model Dev.*, 12, 2463–2480, <https://doi.org/10.5194/gmd-12-2463-2019>, 2019a.
- 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, *Geosci. Model Dev.*, 12, 2463–2480, <https://doi.org/10.5194/gmd-12-2463-2019>, 2019b.
- 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, *Hydrol. Earth Syst. Sci.*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019c.
- 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 Resour. Res.*, 56, e2019WR025975, <https://doi.org/10.1029/2019WR025975>, 2020.
- Kolmogorov, A. N.: Foundations of the theory of probability, *Foundations of the theory of probability*, Chelsea Publishing Co., Oxford, England, 71, p. 8, 1933.
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, *Meteorol. Z.*, 15, 259–263, <https://doi.org/10.1127/0941-2948/2006/0130>, 2006.
- Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., and Wisser, D.: High-resolution mapping of the world's reservoirs and dams

- for sustainable river-flow management, *Front. Ecol. Environ.*, 9, 494–502, <https://doi.org/10.1890/100125>, 2011.
- Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Uijlenhoet, R., Mizukami, N., and Clark, M. P.: HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models, *Hydrol. Earth Syst. Sci.*, 20, 1069–1079, <https://doi.org/10.5194/hess-20-1069-2016>, 2016.
- Messenger, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, *Nat. Commun.*, 7, 13603, <https://doi.org/10.1038/ncomms13603>, 2016.
- 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 Resour. Res.*, 53, 8020–8040, <https://doi.org/10.1002/2017WR020401>, 2017.
- Mott, R., Vionnet, V., and Grünwald, T.: The Seasonal Snow Cover Dynamics: Review on Wind-Driven Coupling Processes, *Front. Earth Sci.*, 6, 197, <https://doi.org/10.3389/feart.2018.00197>, 2018.
- Myneni, R., Knyazikhin, Y., and Park, T.: MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006, MODIS [data set], <https://doi.org/10.5067/MODIS/MCD15A3H.006>, 2015.
- Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I, A discussion of principles, *J. Hydrol.*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.
- Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrol. Earth Syst. Sci.*, 19, 209–223, <https://doi.org/10.5194/hess-19-209-2015>, 2015.
- Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, *Science*, 263, 641–646, <https://doi.org/10.1126/science.263.5147.641>, 1994.
- 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, *J. Hydrol.*, 522, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>, 2015.
- Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., Hagen, J.-O., Hock, R., Kaser, G., Kienholz, C., Miles, E. S., Moholdt, G., Mölg, N., Paul, F., Radić, V., Rastner, P., Raup, B. H., Rich, J., Sharp, M. J., and Consortium, T. R.: The Randolph Glacier Inventory: a globally complete inventory of glaciers, *J. Glaciol.*, 60, 537–552, <https://doi.org/10.3189/2014JoG13J176>, 2014.
- Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, *Hydrol. Sci. J.*, 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, *J. Geophys. Res.-Atmos.*, 124, 13991–14007, <https://doi.org/10.1029/2019JD030767>, 2019.
- Reggiani, P. and Schellekens, J.: Modelling of hydrological responses: the representative elementary watershed approach as an alternative blueprint for watershed modelling, *Hydrol. Process.*, 17, 3785–3789, 2003.
- Reggiani, P., Sivapalan, M., and Majid Hassanizadeh, S.: A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics, *Adv. Water Resour.*, 22, 367–398, [https://doi.org/10.1016/S0309-1708\(98\)00012-8](https://doi.org/10.1016/S0309-1708(98)00012-8), 1998.
- Reggiani, P., Hassanizadeh, S. M., Sivapalan, M., and Gray, W. G.: A unifying framework for watershed thermodynamics: constitutive relationships, *Adv. Water Resour.*, 23, 15–39, [https://doi.org/10.1016/S0309-1708\(99\)00005-6](https://doi.org/10.1016/S0309-1708(99)00005-6), 1999.
- Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, *Geosci. Model Dev.*, 13, 1179–1199, <https://doi.org/10.5194/gmd-13-1179-2020>, 2020.
- Rouholahnejad Freund, E., Zappa, M., and Kirchner, J. W.: Averaging over spatiotemporal heterogeneity substantially biases evapotranspiration rates in a mechanistic large-scale land evaporation model, *Hydrol. Earth Syst. Sci.*, 24, 5015–5025, <https://doi.org/10.5194/hess-24-5015-2020>, 2020.
- Rutter, A. J., Kershaw, K. A., Robins, P. C., and Morton, A. J.: A predictive model of rainfall interception in forests, I. Derivation of the model from observations in a plantation of Corsican pine, *Agr. Meteorol.*, 9, 367–384, [https://doi.org/10.1016/0002-1571\(71\)90034-3](https://doi.org/10.1016/0002-1571(71)90034-3), 1971.
- Rutter, A. J., Morton, A. J., and Robins, P. C.: A Predictive Model of Rainfall Interception in Forests, II. Generalization of the Model and Comparison with Observations in Some Coniferous and Hardwood Stands, *J. Appl. Ecol.*, 12, 367–380, <https://doi.org/10.2307/2401739>, 1975.
- Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resour. Res.*, 46, 1–25, <https://doi.org/10.1029/2008WR007327>, 2010.
- Schaefli, B. and Gupta, H. V.: Do Nash values have value?, *Hydrol. Process.*, 21, 2075–2080, <https://doi.org/10.1002/hyp.6825>, 2007.
- Schellekens, J., van Verseveld, W., Visser, M., hewinsemius, laurenebouaziz, tanjaeuser, sanderdcdevries, cthiange, hboisgon, DirkEilander, DanielTollenaar, aweerts, Baart, F., Pieter9011, Pronk, M., arthur lutz, ctenvelden, Imme1992, and Jansen, M.: openstreams/wflow: Bug fixes and updates for release 2020.1.2, Zenodo [code], <https://doi.org/10.5281/zenodo.4291730>, 2020.
- Seibert, J.: On the need for benchmarks in hydrological modelling, *Hydrol. Process.*, 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, *Hydrol. Process.*, 32, 1120–1125, <https://doi.org/10.1002/hyp.11476>, 2018.
- Shrestha, P., Sulis, M., Simmer, C., and Kollet, S.: Impacts of grid resolution on surface energy fluxes simulated with an integrated surface-groundwater flow model, *Hydrol. Earth Syst. Sci.*, 19, 4317–4326, <https://doi.org/10.5194/hess-19-4317-2015>, 2015.

- Shuai, P., Chen, X., Mital, U., Coon, E. T., and Dwivedi, D.: The effects of spatial and temporal resolution of gridded meteorological forcing on watershed hydrological responses, *Hydrol. Earth Syst. Sci.*, 26, 2245–2276, <https://doi.org/10.5194/hess-26-2245-2022>, 2022.
- Smirnov, N.V.: Estimate of Deviation between Empirical Distribution Functions in Two Independent Samples, *Bulletin Moscow University*, 2, 3–16, 1933.
- Sorooshian, S. and Gupta, V. K.: Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness, *Water Resour. Res.*, 19, 260–268, <https://doi.org/10.1029/WR019i001p00260>, 1983.
- 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., Karssen, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannamete, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, *Geosci. Model Dev.*, 11, 2429–2453, <https://doi.org/10.5194/gmd-11-2429-2018>, 2018.
- Tanaka, T. and Tachikawa, Y.: Testing the applicability of a kinematic wave-based distributed hydrological model in two climatically contrasting catchments, *Hydrol. Sci. J.*, 60, 1361–1373, <https://doi.org/10.1080/02626667.2014.967693>, 2015.
- Tonkin, M. J. and Doherty, J.: A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, 41, W10412, <https://doi.org/10.1029/2005WR003995>, 2005.
- Tromp-van Meerveld, H. J. and McDonnell, J. J.: Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope, *Water Resour. Res.*, 42, W02410, <https://doi.org/10.1029/2004WR003778>, 2006.
- van Verseveld, W., Visser, M., Bootsma, H., Boisgonnier, H., and Bouaziz, L.: Wflow.jl, Zenodo [code], <https://doi.org/10.5281/zenodo.5384924>, 2021.
- Vertessy, R. A. and Elsenbeer, H.: Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization, *Water Resour. Res.*, 35, 2173–2187, <https://doi.org/10.1029/1999WR900051>, 1999.
- Vionnet, V., Marsh, C. B., Menounos, B., Gascoin, S., Wayand, N. E., Shea, J., Mukherjee, K., and Pomeroy, J. W.: Multi-scale snowdrift-permitting modelling of mountain snowpack, *The Cryosphere*, 15, 743–769, <https://doi.org/10.5194/tc-15-743-2021>, 2021.
- Vrugt, J. A., Bouten, W., Gupta, H. V., and Sorooshian, S.: Toward improved identifiability of hydrologic model parameters: The information content of experimental data, *Water Resour. Res.*, 38, 48-1–48-13, <https://doi.org/10.1029/2001WR001118>, 2002.
- Wagener, T. and Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, *J. Hydrol.*, 320, 132–154, <https://doi.org/10.1016/j.jhydrol.2005.07.015>, 2006.
- Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., Adeniyi, K., Andela, B., Arnone, E., Berg, P., Caron, L.-P., Cionni, I., Corti, S., Drost, N., Hunter, A., Lledó, L., Mohr, C. W., Paçal, A., Pérez-Zanón, N., Predoi, V., Sandstad, M., Sillmann, J., Sterl, A., Vegas-Regidor, J., von Hardenberg, J., and Eyring, V.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP, *Geosci. Model Dev.*, 14, 3159–3184, <https://doi.org/10.5194/gmd-14-3159-2021>, 2021.
- 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, *Sci. Data*, 5, 180118, <https://doi.org/10.1038/sdata.2018.118>, 2018.
- Wood, E. F., Sivapalan, M., Beven, K., and Band, L.: Effects of spatial variability and scale with implications to hydrologic modeling, *J. Hydrol.*, 102, 29–47, [https://doi.org/10.1016/0022-1694\(88\)90090-X](https://doi.org/10.1016/0022-1694(88)90090-X), 1988.
- Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, *Water Resour. Res.*, 47, W05301, <https://doi.org/10.1029/2010WR010090>, 2011.
- Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset, *Water Resour. Res.*, 55, 5053–5073, <https://doi.org/10.1029/2019WR024873>, 2019.