Large-sample assessment of ~~varying~~ *varying* spatial ~~scaling effects~~ *resolution* on the *streamflow estimates* of the ~~distributed~~ wflow_sbm hydrological model~~shows that finer spatial resolution does not necessarily lead to better streamflow estimates~~

Jerom P.M. Aerts[1], Rolf W. Hut[1], Nick C. van de Giesen[1], Niels Drost[2], Willem J. van Verseveld[3], Albrecht H. Weerts[4,5], and Pieter Hazenberg[6]

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

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

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

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

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

[6]Applied Research Center, Florida International University, FL 33174, Miami, the United States of America.

Correspondence: Jerom Aerts (J.P.M.Aerts@tudelft.nl)

Abstract. Distributed hydrological modelling moves into the realm of hyper-resolution modelling. This results in a plethora of scaling related challenges that remain unsolved. In light of model result interpretation, finer resolution output might implicate to the user an increase in understanding of the complex interplay of heterogeneity within the hydrological system. Here we investigate spatial scaling in the ~~realm of hyper-resolution~~ *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 ~~were~~ *are* derived at 3 spatial resolutions, namely 3 km, 1km, and 200m. 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 (KGE score) of the 3 model instances ~~show~~ *resulted in* only a statistical difference between the 3km and 200m streamflow estimates. However, ~~results~~ *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 ~~scaling behaviour~~ *the differences* between model instances expressed by differences in KGE scores of on average 0.22 ~~. This demonstrates the presence of scaling behavior throughout the domain and indicates where locality in results is strong.~~ *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 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 among which parameter estimation and the representation of spatial heterogeneity (Clark et al., 2017).

In 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 techniques to overcome the parameter identifiability problem (e.g. Sorooshian and Gupta, 1983; Vr . 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; . 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 technique (MPR) 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 pedo-transfer functions to derive parameters at the finest available data resolution and upscale 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 behavior. This type of research is still very relevant as in recent years, the hydrologic modelling community is making an effort to 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, 19 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 due to 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. An 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). Further

Because of 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. ~~The~~ In the case of spatial model resolution, the increase in level of detail in model output might implicate to the user 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 ~~different spatial and temporal~~ various spatial (and temporal) resolutions.

Multiple studies have ~~tested~~ 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 others effects, that soil moisture estimates were grid cell size dependent. Sutanudjaja et al. (2018) introduced the transition from 30 arc minutes to 5 arc minutes 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 hydrological diverse basins should be considered when investigating ~~spatial scaling effects~~ the effects of varying spatial resolution. To our knowledge there are no studies that have looked into ~~scaling effects~~ varying spatial resolution within the hyper-resolution realm on a large-sample of basins. This will be the focus of this research.

~~Studies have looked in to the scaling of parameters. A study by Benedict et al. (2017) solely increased the spatial resolution of vegetation and orography parameters from 0.5 to 0.05 degrees in a distributed model. Their findings showed no significant improvements in streamflow estimates. They attributed this to hydrological processes depending heavily on other parameters at a coarser resolution. The multi-scale parameter regionalization technique (MPR) introduced by Samaniego et al. (2010) uses transfer functions (e.g. arithmetic mean, harmonic mean) to scale model parameters from the finest available data scale to the hydrological model resolution. A study by Mizukami et al. (2017) demonstrated how the MPR technique can create seamless parameter maps over a large-domain without loss of skill compared to patchwork parameter maps. The method was revisited to estimate parameters of the PCR-GLOBWB model (Samaniego et al., 2017). Results showed the benefits of flux-matching conditions across scales with an increase in model efficiency and better consistency of evapotranspiration fields. Imhoff et al. (2020) applied a flux-matching method to upscale parameters derived at the native data resolution to the model resolution of the wflow_sbm hydrological for the Rhine basin. Similar to Samaniego et al. (2017), this resulted in better consistency of evapotranspiration fields across scales. Simulated streamflow estimates in headwaters were found to be inconsistent across scales. The simulations of the main river Rhine were scale invariant.~~

The distributed conceptual wflow simple bucket model (wflow_sbm) (Schellekens et al., 2020) utilizes high resolution datasets 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 scale. Remotely sensed soil and land cover data sets are sources for estimating

parameters through pedo-transfer functions (PTF) (e.g. Brakensiek et al., 1984; Cosby et al., 1984). 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 is scarce. For most wflow_sbm model parameters a priori parameters are available. ~~No PTF~~ This is not (yet) the case for the horizontal conductivity fraction (KsatHorFrac) ~~is yet available~~ 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 ~~by Imhoff et al. (2020))~~ in Imhoff et al. (2020).

In this study we quantify the effects of ~~spatial scaling~~ varying spatial resolution on the wflow_sbm streamflow estimates for a large-sample of hydrological diverse basins in the CAMELS dataset (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 3 model instances at varying spatial resolutions for each basin: a 3km, 1km, and 200m spatial grid resolution. These instances cover a broad range of large and small scale dynamics. For example: snow accumulation at the mountain range (> 1km) and the mountain ridge scale (< 1km) (e.g. Houze Jr., 2012; Mott et al., 2018; Vionnet et al., 2021), or closing in on the hillslope scale (< 100m) (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. In addition with this study, we showcase how the eWaterCycle platform (Hut et al., 2021) can be utilized for computational intensive large-sample modelling studies.

2 Methodology

2.1 ~~Data~~ 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 hydrological and climatological 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.
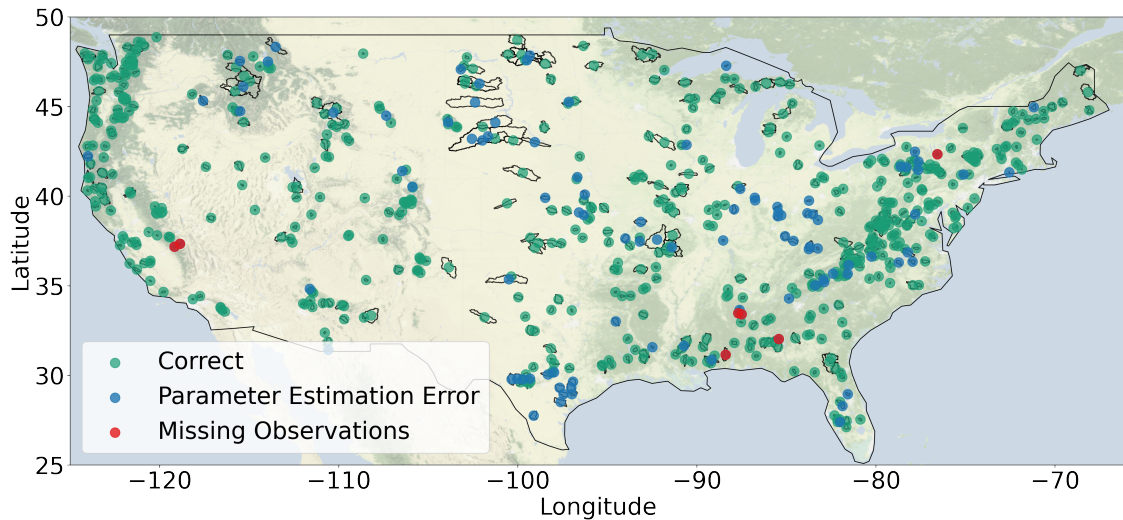
Figure 1. Basin locations of the CAMELS data set with in green the ~~successful model runs~~ included basins. In blue ~~failed runs~~ excluded basins due to parameter ~~or run time~~ estimation errors and in red excluded basins due to missing streamflow observations. Basemap made with Natural Earth.

Of the 671 basins we ran 567 basins successfully for each of the 3 model instances (i.e. 3km, 1 km and 200m resolution)~~successfully for 567 basins . Failed runs were either caused by~~. Reasons for excluding basins in our analyses are missing streamflow observations (7 basins) or errors during parameter ~~derivation, errors during run time, or missing streamflow observations. If~~ estimation (97 basins). Parameter estimation errors occurred mainly during drainage network delineation, either when the basin outlet consisted of a single grid cell that results in a model coding error or when inconsistencies occurred in the local drainage direction layer. When a single model instance of the 3 model instances failed the basin was excluded from further ~~analysis~~ analyses. Figure 1 shows the locations of the ~~successful and failed model runs~~ included and excluded basins as well as the reason for exclusion.

2.1.2 Streamflow Observations

The United States Geological Survey (USGS) streamflow observation records were downloaded to match our model simulation period from 1996 to 2016. The data is resampled to daily data and the units were converted to ~~m3/s~~m3*s-1. We ensured consistency in time zones between the observation and the 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_noteboo

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 was obtained from the Multi-Source Weighted-Ensemble Precipitation Version 2.1 (MSWEP) (Beck et al., 2019). The data set was constructed using bias corrected gauge, satellite and reanalyses data. The data is available at 0.1 degrees spatial (~~~11km~~11 km) and 3-hourly temporal resolution for the period 1979 to 2017. The temperature variable was obtained from the ERA5 reanalyses data set (Hersbach et al., 2020). The data is available at 0.25 degrees (~~~31km~~31 km) spatial and a 1-hourly 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 ~~that indicated that the precipitation~~variable for 6 basins were we compared model simulations based on streamflow estimates that use ERA5 precipitation to those that use MSWEP precipitation. Results indicated that use of ERA5 precipitation did not produce desirable streamflow estimates ~~. Results of this analysis are included in the data repository.~~ compared to MSWEP precipitation. Switching to the MSWEP precipitation product improved streamflow estimates throughout the case study area. Figure A1 and Table A1 in the Appendix A contain the results of this analysis. Noticeable is that 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) V2.0 (Righi et al., 2020; Weigel et al., 2021). Before further processing the data is aggregated 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 subsequently used to calculate potential evapotranspiration. The code for ~~all these~~ used for pre-processing ~~steps~~ is included in the Jupyter Notebooks made available with this manuscript (DOI~~:~~: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, landcover & landuse, soil, meteo, and river gauge data. The pedo-transfer function to estimate soil properties is based on

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	Messager et al. (2016)	1.0
Randolph Glacier Inverntory	surface water	Pfeffer et al. (2014)	6.0
CHELSA	meteo	Karger et al. (2017)	1.2
Köppen-Geiger	meteo	Kottek et al. (2006)	2017
VITO	landuse & landcover	Buchhorn et al. (2020)	v2.0.2
Modis LAI	landuse & landcover	Myneni et al. (2015)	MCD15A3H V006
SoilGrids	soil	Hengl et al. (2017)	2017

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 km2 and 10 km2 respectively.

160   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)

~~Wflow_sbm~~ The wflow_sbm model is available as part of the wflow ~~open source~~ open-source modeling framework (Schellekens et al., 2020), which is based on PCRaster (Karssenberg et al., 2010) and Python. Figure 2 shows the different processes and

165   fluxes that are part of the wflow_sbm hydrological concept. The soil part of wflow_sbm model is largely based on the Topog_SBM model (Vertessy and Elsenbeer, 1999), that considers 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, 1995; Ciarapica and Todini, 2002), G2G (Bell et al., 2007), 1K-DHM (Tanaka and Tachikawa, 2015) and Topog_SBM (Vertessy and Elsenbeer, 1999). Wflow_sbm has a simplified phys-

170   ical basis with parameters that represent physical characteristics, leading to (theoretically) an easy linkage of the parameters to actual physical properties. Topog_SBM is mainly used to simulate fast runoff processes during discrete storm events in small ~~catchments~~ 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 ~~catchments~~ basins. The main differences of wflow_sbm with Topog_SBM are:

175   – The addition of evapotranspiration and interception losses (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).
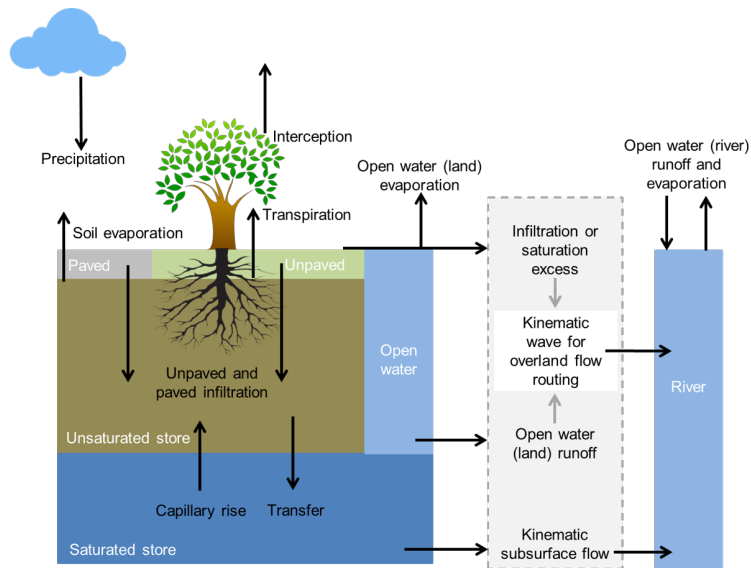
Figure 2. Overview of the different processes and fluxes in the wflow_sbm model (Schellekens et al., 2020).

- The addition of capillary rise.

- The addition of glacier and snow build-up and melting processes.

- Wflow_sbm 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.

- Vertical transfer of water 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 ~~Instances and Parameter Estimation~~ Runs & Calibration

We derived 3 model instances at varying spatial model resolution that cover a 3km, 1km, and 200m grid. ~~The parameter sets 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, landcover & landuse, soil, meteo, and river gauge data. The PTF to estimate soil properties is based on Brakensiek et al. (1984). An overview of the data and references are provided in 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 Messager et al. (2016) 1.0 Randolph Glacier Inverntory surface water Pfeffer et al. (2014)~~

2.2.3 ~~Model Runs & Calibration~~

While for most parameters of the wflow_sbm model a priori estimates can be derived from external sources~~as explained above~~, 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. ~~To do this calibration the model simulations (3km, 1km, and 200m)~~
200 ~~are divided into 2 periods; a calibration period from 1997 to 2006 and an evaluation period from 2008 to 2016. The years 1996 and 2007 are regarded as spin-up years and not included in the analysis of the results. We assessed the streamflow estimates of both periods using the Kling-Gupta Efficiency score (KGE) (Gupta et al., 2009).~~

~~The calibration procedure finds an optimal parameter value based on the KGE objective function of streamflow estimates at the basin outlet. A single non-distributed parameter was calibrated, the saturated horizontal conductivity fraction (KsatHorFrac).~~
205 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). ~~Increasing~~ A sensitivity analyses conducted by the model developers concluded that the 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 ~~reduces~~ reduced peak flow and flashiness.
210 We calibrated the models to match model setups of those used by the users of the hydrological model.

2.2.3 ~~eWaterCycle platform~~

~~This research was conducted within the eWaterCycle platform (Hut et al., 2021). eWaterCycle follows by design the FAIR principles of data science (Wilkinson et al., 2018) and allows high level communication with models regardless of programming language through the Basic Modelling Interface (Hutton et al., 2020). This study showcases how eWaterCycle handles the~~
215 ~~setup of extensive modelling studies. A Jupyter Notebook with the model experiments of this study is provided in the GitHub repository. As 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 a HPC. The calibration~~ 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. 21 runs are evaluated based on an
220 interval of KsatHorFrac values ranging between 1 and ~~evaluation procedures totalled 41.025 model runs on the Dutch national super-computer Cartesius hosted by Surf~~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 analyses.

2.3 ~~Analyses of results~~Benchmark Selection

2.3.1 ~~Benchmark Selection~~

To select basins ~~where model performance is~~ with model performance considered to be at least reasonably well, 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; Schaefli 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. 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~~ 3 resolutions, has a lower KGE than either the median or mean flow benchmark, it is considered not suited for this study and removed from ~~the list of basins.~~ further analyses.

2.4 Analyses of results

2.4.1 Objective function and sampling uncertainty

We use the modified Kling-Gupta Efficiency metric for the analyses 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 KGE score is defined by its 3 components, the Pearson correlation, the mean bias, and the variability bias. The lowest value of the 3 components determines the KGE score. The KGE score and its components:

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

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

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

,with r the Pearson correlation, β the mean bias, and γ the variability bias. μ_s and σ_s are the mean and standard deviation of the simulations. 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 jacknife-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 the 95th percentiles. We ran the gumboot analyses 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~~ streamflow estimates

To provide more context to 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 study by Knoben et al. (2020) ran 36 conceptual models using the Modular Assessment of Rainfall-Runoff Models Toolbox V1.0 (MARRMoT) (Knoben et al., 2019a, b) on the CAMELS data set. 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 directly compared to those of this study. It does however provide context to 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 Kolmogarov-Smirnoff test (Kolmogorov, 1933; Smirnov, N.V., 1933) to test if the differences between the KGE score distributions of the model instances are statistically relevant. ~~Allowing~~ 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). eWaterCycle follows by design the FAIR principles of data science (Wilkinson et al., 2018) and allows high level communication with models regardless of programming language through the Basic Modelling Interface (Hutton et al., 2020). This study showcases how eWaterCycle handles the setup of extensive modelling studies. A Jupyter Notebook with the model experiments of this study is provided in the GitHub repository. As 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 a HPC. The calibration and evaluation procedures totalled 41.025 model runs on the Dutch national super-computer Cartesius hosted by Surf.

3 Results

The results in this section are based on the ~~KGE objective function~~ modified KGE 2012 objective function applied to the streamflow estimates at the basin outlet~~for the evaluation period. Results based on the~~. The Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) results are available in the repository (DOI~~:~~:10.5281~~/~~/zenodo.5724576). ~~The first part describes the effect of spatial scaling on model calibration, model evaluation and terrain characteristics. The second part describes the streamflow analyses results for the CAMELS data set.~~

3.1 ~~The effect of calibration on streamflow estimates~~Calibration period results

3.1.1 The effect of calibration on streamflow estimates

To illustrate how model calibration effects the streamflow estimates of each model instance, we first show the calibration curves of a single basin (ID:14301000). To avoid presentation bias we selected a basin with moderate performance and only show the last year of calibration. Figures 3abc show 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. Not all calibration interval values are shown in the figure for visualization purposes. The results depict the effect that increasing the KsatHorFrac values has on the hydrograph: base flow increases whilst peak flow reduces. In addition, large KsatHorFrac values reduce the flashiness of the streamflow estimates as is visible in Figures 3b and 3c in the second week of November 2005. Of note is that the selection of best calibration parameter values is strongly dependent on the chosen objective function as for example the NSE score would be more favourable towards flashiness and less towards base flow. As shown in Figure 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.

3.2 ~~The effect of spatial scale on terrain characteristics~~

~~We illustrate the effect of spatial scaling on the parameter set~~ Figure 4 shows the CDFs of the KGE score distribution based on the best performing calibration model run of the 3 model instances~~by showing the difference in topography and drainage density for 3 basin. 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 probability density function (PDF) of the height distributions of~~. Starting with the modified KGE score distributions in Figure 4a we find that the model instance distributions are very similar for the ~~model instances for each of the 3 basins, Figure 11bshows the slope distribution, and Figure 11cthe profile curvature distributions.~~

~~The height distribution of the DEM in Figure 11a shows , most clearly for basin ID 02231342, how the representation of the highest altitudes is underestimated by the 3km model instance (orange) compared to the~~ whole domain, especially the 1km and 200m ~~instance (green) and to a lesser extent the 1km 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 Figures 11bc. As can be expected from the height distributions, the slope of~~ instances. The 3km instance has lower scores for 60 % of the distribution 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 (Figure 4b) and gamma variability ratio (Figure 4c). The largest differences are visible in the beta bias component (Figure 4d that shows similar bias for the ~~200m instancehas more gentle and steep sloping topography than the~~ 1km and 200m with 60 % of the distribution lower than a value of 1 and 40 % higher. With the highest agreement in the lower and upper 5% of the
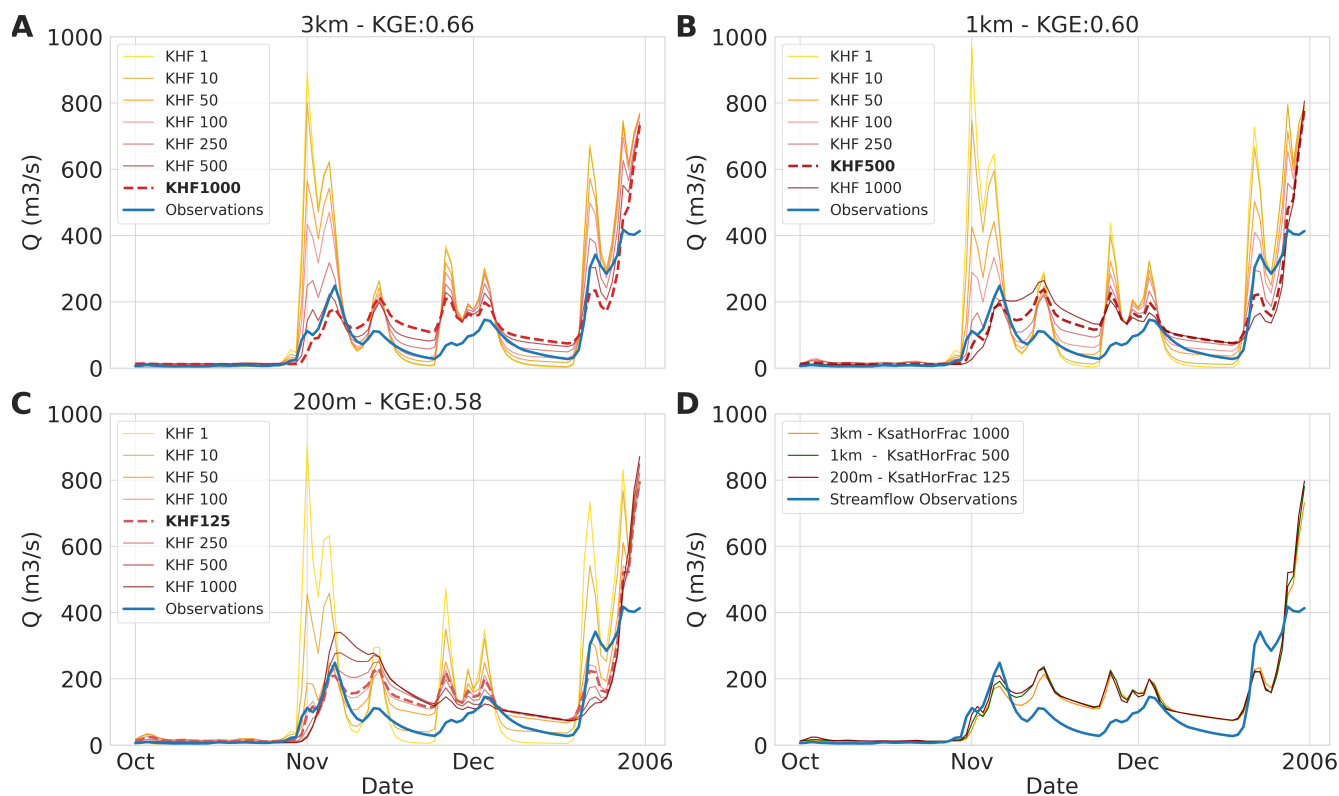
Figure 3. Calibration period results. The calibration interval of the KsatHorFrac parameter (KHF) for the 3 model instances at the basin outlet (ID:14301000) of the final calibration period year. (a) 3km model instance. (b) 1km model instance. (c) 200m model instance. Values range from 1 to 1000 (yellow to red), 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 3km (orange), 1km (green), and 200m (red) model instances.

distribution. The 3km ~~instances~~instance agrees only in the upper 10% of the distribution with the 1km and 200m instances with 70 % of the 3km distribution lower than a value of 1 and 30 % 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 the 20 % of the distributions with larger than 2.0 bias values.

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 of calendar day, either mean or median during the model evaluation period. Figure 5a shows spatially which type of calendar day benchmark is best performing per basin. Of the 567 simulated basins the results of 454 basins exceed the benchmark
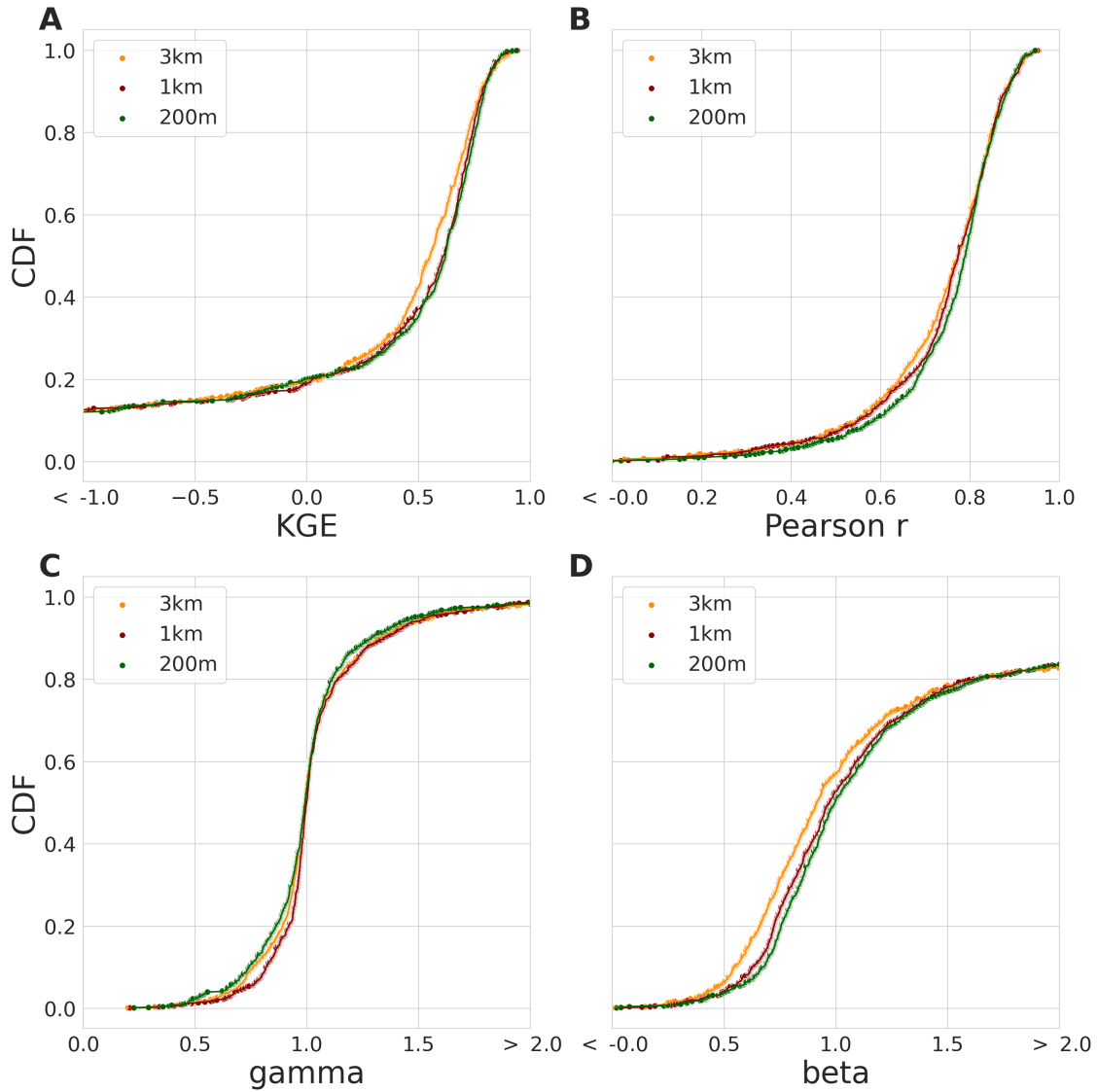
Figure 4. Calibration period results. The modified KGE score CDFs of the best performing model runs during the calibration period and its individual components of 3 model instances. With in orange the 3km, in red the 1km, and in green the 200m model instances. (a) The CDF of the modified KGE score. (b) 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.

for each model instance. 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 Figure 11c indicates whether a slope is linear (values close to 0), concave (values smaller than 0), and convex (values larger than 0). The 3km and 1km instances show similar slope geometries. 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 KGE 2012 score distribution of the best performing benchmark type. Basemaps are made with Natural Earth.

325   ~~the 3km instance having slightly more linear slopes. At the finest resolution (200m) the slopes geometry shifts from linear slopes to either convex or concave curvature profiles.~~ 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 Figure 5c and the distribution of the benchmark KGE scores is shown in Figure 5d. The distribution ranges from -0.62 to 0.71 KGE score and is

330   skewed towards values lower than 0.0 with a mean of 0.02 and median of -0.02.

~~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~~

The statistical benchmarks (mean and median during the evaluation period) are plotted as CDFs based on the KGE score and its 3 individual components in Figure 6. For most of the KGE distributions in Figure 6a the mean benchmark outperforms

335   the median with the upper 10 % being the exception. The Pearson correlation coefficient (Figure 6b) is slightly higher for the

median at 70 % of the distribution. In the case of the gamma variability ratio (Figure 6c) the bottom 38 % is lower for the median than the mean and the upper 55 % slightly higher for the median than the mean. The determining component of the KGE score is the beta component shown in Figure 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, the lower 10% and the upper 1% of the distribution. These basins have flow regimes that greatly differ per year compared to the climatology. Considering 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 out of the 44 basins.

3.3 ~~The effect of spatial scale on streamflow estimates~~

3.2.1 The effect of spatial resolution on streamflow estimates

The same 3 example basins as with the calibration period 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 Figure 7.

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

3.3 ~~Benchmark selection~~

~~We calculate a statistical benchmark to determine basins where the streamflow estimates of the model instances are deemed adequate for further analyses. The best performing type of climatology of calendar day benchmark, either mean or median, is depicted in Figure 5a. Figure 5b shows which basins are accepted and which basins are rejected based on the 10 year climatological benchmark. 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 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 Figure 5c and the distribution of the benchmark KGE scores is shown in Figure 5d. The distribution ranges from -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. A KGE score of -0.41 is comparable to taking the mean flow during the evaluation period (Knoben et al., 2019).~~

~~(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~~
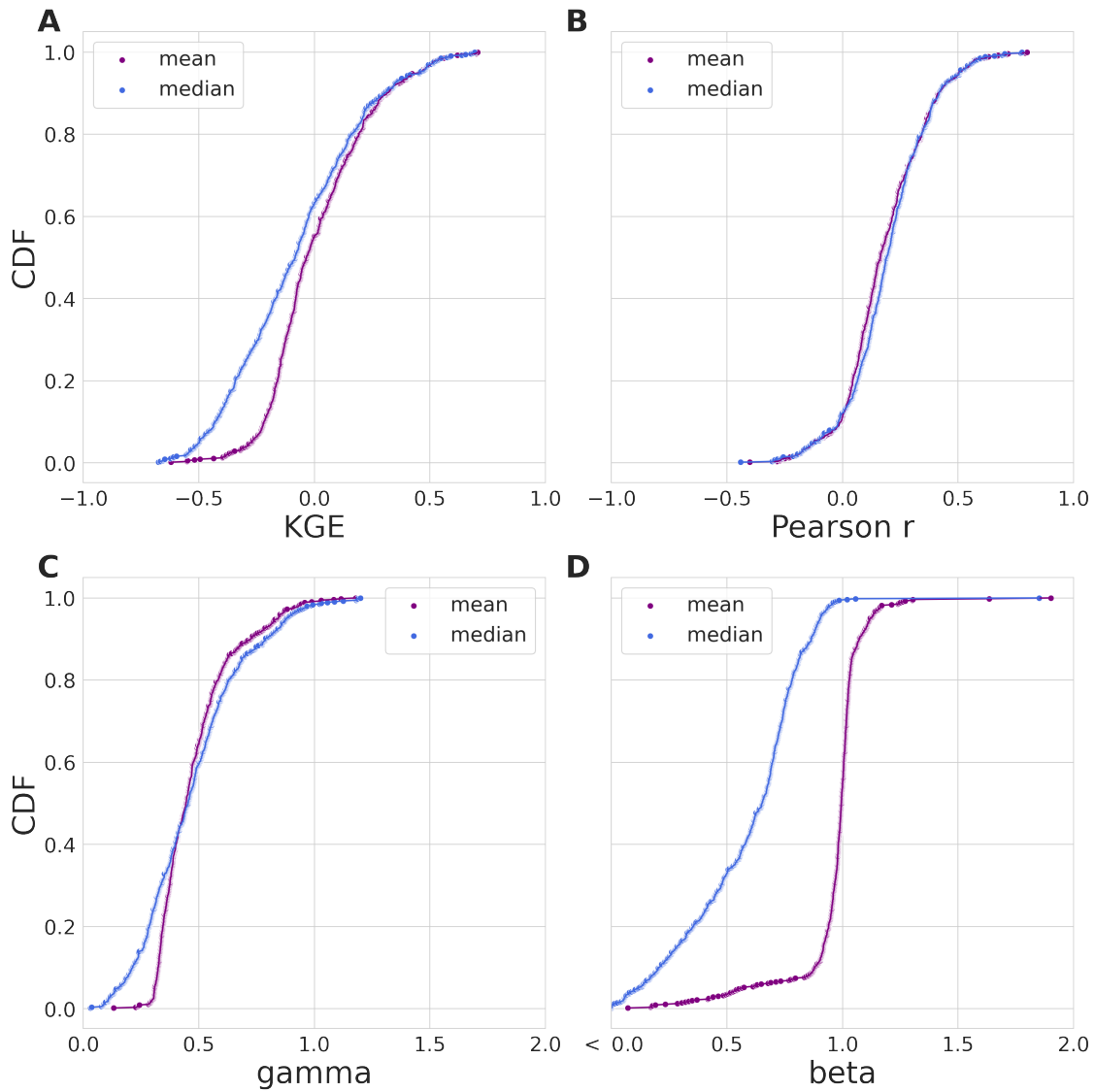
Figure 6. ~~Evaluation period results.~~ The ~~drainage density defined as stream length divided basin area for~~ CDFs of the modified KGE score ~~and its 3~~ ~~example basins~~ individual components of the statistical benchmark during the evaluation period. With in purple the mean and in blue the median statistical benchmark. (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.

~~0.0133 0.0141 0.0143 0.0056 0.0059 0.0063 0.0112 0.0123 0.0126~~

~~are accepted or rejected and the best performing benchmark type. (d) The KGE score distribution of the best performing benchmark type. Basemaps made with Natural Earth.~~

3.3 ~~Streamflow estimates of model instances~~

~~The KGE score~~

3.2.1 Streamflow estimates of model instances

The KGE score results for the evaluation period are shown in Figure ~~5 in the form of an Cumulative Distribution Function (CDF)~~8a. The KGE score distribution of the mean of 36 hydrological models from ~~Knoben et al. (2020)~~ Knoben et al. (2020) is included and is referred to as "MARRMoT mean results". Of note is that 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 (Figure ~~5~~8, blue) of ~~Knoben et al. (2020)~~ Knoben et al. (2020) is close to the mean of the distributions of the 3 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 KGE score distribution is higher than the 1km model instance. The MARRMoT mean results for the lower 5 % of the CDF performs better than the 3 model instance distributions. Here, the range of KGE scores is smaller for the MARRMoT mean (-1.55 to 0.09) than for the 3 model instances (-13.56 to 0.00). Above 0.17 of the CDF (83 % of the results) the distributions of the 3 model instances are higher in KGE scores than those of the MARRMoT mean.

When we consider only 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 3km instance has the highest and the 1km the lowest KGE score. At 40 % of the distribution and lower the 200m instance is followed by the 1km and 3km instances in terms of highest KGE score. The Pearson r component of the KGE score CDF in Figure 8b and the gamma variability ratio component 8c show small differences between model instances. The beta bias component in Figure 8d shows the largest differences between model instances. Especially between the 3km and the 1km and 200m model instances. The bias component is the main factor for the differences in the overall KGE score.

Next, we apply the Kolmogorov-Smirnov (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 3km and 200m model instances is statistical relevant at a p-value of ~~0.01~~0.02. On a large-sample, this means that increasing the spatial model resolution from 3km to 1km or 1km to 200m does not lead to significant differences in streamflow performance. ~~However, when~~ When changing resolution from 3km to 200m, the distribution of KGE scores is significantly different ($p < 0.05$) according to the KS-test.

~~The CDF~~

3.2.2 Objective function uncertainty

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

CDFs	Kolmogorov-Smirnov Statistic	p-value
3 km - 1 km	~~0.07~~ 0.08	~~0.21~~ 0.14
1 km - 200 m	~~0.06~~ 0.05	~~0.31~~ 0.82
3 km - 200 m	~~0.10~~ 0.11	~~0.01~~ 0.02

Table 3. Sample uncertainty analyses results per quarter of the total percentage of the modified KGE cumulative distribution function of the evaluation period. The mean of the 3 model instance 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 2x std err Jackknife	Mean 2x 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

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 is shown in Figure 9.The results for the 3 model instances are very similar. The tolerance interval results denoted with the black lines show that approximately a 100 basins have a KGE sampling uncertainty of 0.1 or lower and approximately half of the basins 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 8 by calculating the mean of the tolerance interval, jackknife standard error, and bootstrap standard error for each of the 3 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 the lowest average values.Indicating that sample uncertainty is high at low KGE values and vice versa.

3.2.3 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, Figure 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 Figure 10a, are found in the Pacific Northwest, Atlantic South, Appalachia, and Northeast of the CONUS. KGE scores lower than -0.41 are found throughout the CONUS. The highest KGE scores in Figure 10b are located in the Northwest, Rocky Mountains, and Appalachia. These regions are characterised 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 South, and the Midwest. These regions span a wide range of hydro-climatic diverse basins. The average KGE score difference is ~~0.78~~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 instance there is no overall geographical trend in results. Best performances for the 1km model instance are generally close to basins where the 200m model instance is best performing. The 3km model instance shows some clusters in the South and Pacific Northwest. The Rocky Mountains contains the best performing 200m model instances and the Appalachia a mixture between 1km and 200m 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 3 model instances by showing the difference in topography and drainage density for 3 basin. 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 probability density function (PDF) of the height distributions of the model instances for each of the 3 basins, Figure 11b shows the slope distribution, and Figure 11c the profile curvature distributions.

The height distribution of the DEM in Figure 11a shows, most clearly for basin ID 02231342, how the representation of the highest altitudes is underestimated by the 3km model instance (orange) compared to the 200m instance (green) and to a lesser extent the 1km 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 Figures 11bc. As can be expected from the height distributions, the slope of the 200m instance has more gentle and steep sloping topography than the 1km and 3km 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 Figure 11c indicates whether a slope is linear (values close to 0), concave (values smaller than 0), and convex (values larger than 0). The 3km and 1km instances show similar slope geometries. With the 3km instance having slightly more linear slopes. At the finest resolution (200m) the slopes geometry shifts from linear slopes to either convex or concave curvature profiles.

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 3 basins.

4   Discussion

4.1 Benchmarks

We applied an initial statistical benchmark based on streamflow observations for basin selection to identify basins ~~where~~ which streamflow estimates are deemed adequate for further analyses. This does not imply that excluded basins are less relevant.

Table 4. The drainage density defined as stream length divided basin area for the 3 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

450 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 CDFs of the KGE score and its components based on the statistical benchmarks in Figure 6 show that 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 not close to 1 for the mean benchmark indicate that the flow regime changes from year to year. The hydrological

455 model simulation were in most cases 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).~~ (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

460 of 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 the CONUS as case study area (e.g. Mizukami et al., 2017; Rakovec et

465 . 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

470 further propose the use of a model experiment environment, such as eWaterCycle ~~(Hut et al., 2021),~~ (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 (DOI~~:~~: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 deemed not

475 suitable for publication that 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 ~~plotted as an CDF (Figure 5)~~ of the calibration period in Figure 4 and the evaluation period in Figure 8 show that this is the case for the KGE score distributions

480   based on streamflow estimates at the basin outlet. ~~The~~ Although the differences are small, the crossing of the distribution lines is a strong indication that ~~their~~ there is disagreement on KGE scores between model instances and that there is no single instance outperforming the rest consistently. ~~For some basins the coarser~~ The largest difference between distribution for both periods is found in the beta bias component of the KGE score. The benchmark selection excluded mainly the basins that contained large bias values during the calibration period. For the bias component, the 3km instance ~~outperformed the finer~~ deviates from the

485   1km and 200m ~~instance, showing that finer resolution does not automatically lead to improved streamflow estimates.~~
~~Next, we applied the~~ instances. This shows that the total volume of streamflow does differ for the 3km instance. This difference was confirmed by testing the statistical significance of the differences between distributions of the model instances by applying the KS-statistic ~~to test if there is a significant difference between the distributions of~~ on the evaluation period results. Here a statistical difference between distributions is only found for model ~~instances. This is only the case for the model~~

490   instance combination 3km and 200m ~~. When we consider the increase in streamflow based model performance as opposed to computational cost, we find that this does not scale linearly with~~ 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 periods. To provide the reader with more context we have included the KGE 2009 (Gupta et al., 2009), modified KGE , KGE non-parametric (Pool et al., 2018), and NSE objective function results in

495   the data repository and in Figure A2. The conclusions based hypothesis testing 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 and observed streamflow and less influenced by 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 for the whole period (Fowler et al., 2018). In addition, as stated in Clark et al. (2021) it is important to determine the

500   sampling uncertainty of objective functions to avoid wrong conclusions at the system scale. Following their methodology we investigated the sampling uncertainty by applying bootstrap and jackknife-after-bootstrap methods (Figure 9). Approximately half of the basins has a KGE uncertainty of 0.2 or lower based on the tolerance interval and half of the basins higher than 0.2 with values surpassing 0.5 KGE uncertainty. The lower 25 % of the cumulative distribution function (Figure 8 contains the highest sample uncertainty and the ~~amount of grid cells in the basin due to lateral connections in the hydrological model.~~

505   ~~The average non-parallel run time of the 3km instance is 157 seconds and that of the 200m instance 12120 seconds with an average number of grid cell difference of 28872 cells. These results point toward the importance of conducting an initial spatial model resolution assessment at the start of~~ 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. More so, when taking into account the model uncertainty and the observation uncertainty. This demonstrates the drawbacks

510 of conducting large-sample ~~assessments as it avoids subpar 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 improvement (threading, mpi)are being implemented.~~ assessments and the sensitivity to sampling uncertainty of the results.

4.3 Relative model instance differences

515 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 ~~scaling issues~~ variations due to spatial resolution are in effect throughout the ~~spatial~~ domain. This is depicted by the differences between the KGE scores of the model instances (Figure 10c). On average these is a 0.22 KGE score difference with extremes of more than 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) the

520 only form of validation that is actually possible. Therefore, we argue that the differences are large locally even though they might be within the sampling uncertainty range.

We find that the 1km instance is best performing in basins where the difference between minimum and maximum KGE score of the 3 instances is small (Figure 10c). We attribute this partly to the calibration routine finding a better optimal parameter value (KsatHorFrac) as results are often close to those of the 200m instance. For the best performing 3km or 200m model

525 instances (Figure 10d) there are only small geographical trends of best model performance in the South and Appalachia. This information is valuable for future research that does a more in-depth assessment of the internal states and fluxes as we now know where the effects of varying grid cell size is small or large and which model is best performing.

We conducted a terrain analyses (Figure 11) to identify changes in terrain characteristics due to spatial resolution that might explain the differences between model instance streamflow estimates. Minor effects are depicted by the profile curvatures

530 present at 200m 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 up-scaling method of Eilander et al. (2021) is (almost) consistent across spatial scale.

Larger differences between model instances are found for the height distribution of the DEM, which is flattened at course

535 resolution compared to finer resolution. At high altitudes this introduces changes in snow dynamics due to the use of the temperature degree-day method by the hydrological model. The resulting effect on streamflow estimates depend on the relative contribution of snow melt. 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 as 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

540 in a broader distribution of slopes that effects 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 ~~downscaling~~ pre-processing routine for each model instance. This ensured that ~~for example~~ 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 therefore 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). In addition, 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 by comparing the same forcing product with native implementations at various spatial resolutions that 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.

~~The results~~

4.3.1 Computational cost

When we consider the increase in streamflow based model performance as opposed to computational cost, we find that this 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 3km instance is 157 seconds and that of the 200m instance 12120 seconds with an average number of grid cell difference of 28872 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 subpar 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 improvement (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 for example better capture topographic details and are less influenced by native forcing resolution ((Gharari et al., 2020)).

4.4 Outlook

The results from this study help model developers with model refinement by providing them with understanding where and under what circumstances difference 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 output 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.

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

the whole period (Fowler et al., 2018). In addition, as stated in Clark et al. (2021) it is important to determine the uncertainty of objective functions. The effect of uncertainty is kept at a minimum in this study as only a single parameter is calibrated. We selected the KGE 2012 objective function as it is less influenced by extreme combinations of simulated and observed streamflow and less influenced by structural error in the meteorological input. To provide the reader with more context we have included the KGE 2009 (Gupta et al., 2009), KGE 2012 , KGE non-parametric (Pool et al., 2018), and NSE objective function results in the data repository and supplementary. The conclusions based hypothesis testing were not affected by the type of KGE objective function.

We conducted this study as an initial assessment to follow-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. However, we We suggest that future work investigates the basins that show large differences in lateral fluxes and poor model performance. By including or small differences in model performance, lateral fluxes, and effect 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)conclusions can ., see Guse et al. (2021) for a recent overview. Conclusions might then be made on whether increasing spatial resolution leads to increased model fidelity .or not. This should be tested using multiple forcing data sets to test 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. Future research would also benefit from including basin and climate characteristics of the CAMELS data set in order to find statistical relationships explaining locality in results.

5 Conclusions

The aim of this study was to analyse the effects of spatial scaling 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 (3km, 1km, 200m) is only statistically significantly different between the 3 km and 200 m model instances (P<0.05). This confirms at an aggregated level the hypothesis that differences between model instances are small due to quasi-scale invariant parameter set and process descriptions that remain constant across scale 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 to small to base conclusions on.

Table A1. Evaluation period results. Comparison of ~~evalution~~ evaluation period objective function results of the 3km 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	3km	0.20	0.19	0.67	0.65
02297155	3km	0.23	0.04	0.34	0.06
05408000	3km	0.35	-0.03	0.60	0.39
06406000	3km	-1.69	-6.19	0.21	0.15
07057500	3km	0.68	0.53	0.56	0.35
08079600	3km	-5.65	-8.13	-.023	-0.49

– Results show large differences in maximum and minimum KGE scores with an average of 0.22 between model instances throughout the CONUS. Providing 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 answered where locality in results are strong due to ~~spatial scaling effects~~ 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 what conditions locality in spatial scaling related issues occur.

Code and data availability. The software that supplements this study is available at: https://github.com/jeromaerts/eWaterCycle_example_notebooks. The data that supplements this publications is avaialble at DOI:10.5281/zenodo.5724576

Appendix A

A1 ERA5 and MSWEP precipitation forcing comparison

A2 CDFs of multiple objective functions

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

Competing interests. The authors declare that no competing interests are present.

Acknowledgements. We would like to thank the anonymous reviewer and Shervan Gharari for their valuable feedback that helped improve this manuscript. This work has received funding from the Netherlands eScience Center under file number 027.017.F0. We like to thank the RSEs at NLeSC who have co-build the eWaterCycle platform and Surf for providing computing infrastructure.

References

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

Bak, P.: Complexity and Criticality, in: How Nature Works: the science of self-organized criticality, edited by Bak, P., pp. 1–32, Springer, New York, NY, https://doi.org/10.1007/978-1-4757-5426-1_1, 1996.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. v., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, Bulletin of the American Meteorological Society, 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 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, Hydrology and Earth System Sciences, 11, 532–549, https://doi.org/10.5194/hess-11-532-2007, publisher: Copernicus GmbH, 2007.

Benedict, I., van Heerwaarden, C. C., Weerts, A. H., and Hazeleger, W.: An evaluation of the importance of spatial resolution in a global climate and hydrological model based on the Rhine and Mississippi basin, Hydrology and Earth System Sciences Discussions, pp. 1–28, https://doi.org/10.5194/hess-2017-473, publisher: Copernicus GmbH, 2017.

Benning, R.: Towards a new lumped parameterization at catchment scale., Master Thesis, University of Wageningen., 1995.

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 Resources Research, 48, https://doi.org/10.1029/2011WR010982, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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?, Hydrological Processes, 29, 310–320, https://doi.org/10.1002/hyp.10391, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.10391, 2015.

Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: A review, Hydrological Processes, 9, 251–290, https://doi.org/10.1002/hyp.3360090305, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Hydrology and Earth System Sciences, 10, 925–936, https://doi.org/10.5194/hess-10-925-2006, publisher: Copernicus GmbH, 2006.

Booij, M.: Impact of climate change on river flooding assessed with different spatial model resolutions, Journal of Hydrology, 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., ASAE Paper No. PNR-84-203. American Society of Agricultural Engineers, St. Joseph, MI, USA., 1984.

Bras, R. L.: Complexity and organization in hydrology: A personal view, Water Resources Research, 51, 6532–6548, https://doi.org/10.1002/2015WR016958, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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.

Bruin, H. A. R. d., 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, Journal of Hydrometeorology, 17, 1373–1382, https://doi.org/10.1175/JHM-D-15-0006.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2016.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., and Smets, B.: Copernicus Global Land Cover Layers—Collection 2, Remote Sensing, 12, 1044, https://doi.org/10.3390/rs12061044, number: 6 Publisher: Multidisciplinary Digital Publishing Institute, 2020.

Ciarapica, L. and Todini, E.: TOPKAPI: a model for the representation of the rainfall-runoff process at different scales, Hydrological Processes, 16, 207–229, https://doi.org/10.1002/hyp.342, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Hydrology and Earth System Sciences, 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, publisher: Copernicus GmbH, 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 Resources Research, 57, e2020WR029 001, https://doi.org/10.1029/2020WR029001, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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 Resources Research, 20, 682–690, https://doi.org/10.1029/WR020i006p00682, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/WR020i006p00682, 1984.

Efron, B.: Jackknife-After-Bootstrap Standard Errors and Influence Functions, Journal of the Royal Statistical Society: Series B (Methodological), 54, 83–111, https://doi.org/10.1111/j.2517-6161.1992.tb01866.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Statistical Science, 1, 54–75, https://doi.org/10.1214/ss/1177013815, publisher: Institute of Mathematical Statistics, 1986.

Eilander, D. and Boisgontier, H.: HydroMT, 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, Hydrology and Earth System Sciences, 25, 5287–5313, https://doi.org/10.5194/hess-25-5287-2021, publisher: Copernicus GmbH, 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 Resources Research, 55, 1737–1772, https://doi.org/10.1029/2018WR023903, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR023903, 2019.

Feddes, R. A., Kowalik, P. J., and Zaradny, H.: Simulation of Field Water Use and Crop Yield, Wiley, 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, Journal of Hydrology, 203, 21–38, 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 Resources Research, 54, 3392–3408, https://doi.org/10.1029/2017WR022466, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2017WR022466, 2018.

705 Garrick, M., Cunnane, C., and Nash, J. E.: A criterion of efficiency for rainfall-runoff models, Journal of Hydrology, 36, 375–381, https://doi.org/10.1016/0022-1694(78)90155-5, 1978.

Gash, J. H. C.: An analytical model of rainfall interception by forests, Quarterly Journal of the Royal Meteorological Society, 105, 43–55, https://doi.org/10.1002/qj.49710544304, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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
710 Long Short-Term Memory network, Hydrology and Earth System Sciences, 25, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021, publisher: Copernicus GmbH, 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, Hydrology and Earth System Sciences, 24, 5953–5971, https://doi.org/10.5194/hess-24-5953-2020, publisher: Copernicus GmbH, 2020.

715 Grayson, R. B., Moore, I. D., and McMahon, T. A.: Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes, Water Resources Research, 28, 2639–2658, https://doi.org/10.1029/92WR01258, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, publisher:
720 Elsevier, 2009.

Gupta, V. K., Rodríguez-Iturbe, I., and Wood, E. F.: Scale Problems in Hydrology, Water Science and Technology Library, 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, Hydrological Sciences Journal, 65, 1158–1172, https://doi.org/10.1080/02626667.2020.1734204, publisher: Taylor & Francis
725 _eprint: 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 Resources Research, 57, e2021WR030 661, https://doi.org/10.1029/2021WR030661, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021WR030661, 2021.

Haddeland, I., Lettenmaier, D. P., and Skaugen, T.: Reconciling Simulated Moisture Fluxes Resulting from Alternate Hydrologic Model Time
730 Steps and Energy Budget Closure Assumptions, Journal of Hydrometeorology, 7, 355–370, https://doi.org/10.1175/JHM496.1, 2006.

Hengl, T., Jesus, J. M. d., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, 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, e0169 748, https://doi.org/10.1371/journal.pone.0169748, publisher: Public Library of Science, 2017.

735 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., Vil-

laume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803, 2020.

Horritt, M. S. and Bates, P. D.: Effects of spatial resolution on a raster based model of flood flow, Journal of Hydrology, 253, 239–249, https://doi.org/10.1016/S0022-1694(01)00490-5, 2001.

Houze Jr., R. A.: Orographic effects on precipitating clouds, Reviews of Geophysics, 50, https://doi.org/10.1029/2011RG000365, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2011RG000365, 2012.

Hrachowitz, M. and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, Hydrology and Earth System Sciences, 21, 3953–3973, https://doi.org/10.5194/hess-21-3953-2017, publisher: Copernicus GmbH, 2017.

Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., Aerts, J., Albers, T., Alidoost, F., Andela, B., Camphuijsen, J., Dzigan, Y., van Haren, R., Hutton, E., Kalverla, P., van Meersbergen, M., van den Oord, G., Pelupessy, I., Smeets, S., Verhoeven, S., de Vos, M., and Weel, B.: The eWaterCycle platform for Open and FAIR Hydrological collaboration, Geoscientific Model Development Discussions, pp. 1–31, https://doi.org/10.5194/gmd-2021-344, publisher: Copernicus GmbH, 2021.

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 Resources Research, 56, e2019WR026 807, https://doi.org/10.1029/2019WR026807, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR026807, 2020.

Karger, D. N., Conrad, O., Böhner, 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, 170 122, https://doi.org/10.1038/sdata.2017.122, bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Atmospheric science;Biogeography;Hydrology Subject_term_id: atmospheric-science;biogeography;hydrology, 2017.

Karssenberg, 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, Environmental Modelling and Software, 25, 489–502, https://doi.org/10.1016/j.envsoft.2009.10.004, accepted: 2019-07-30T16:07:37Z, 2010.

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

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, publisher: Copernicus GmbH, 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) v1.0: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous space-state formulations, preprint, Hydrology, https://doi.org/10.5194/gmd-2018-332, 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, Hydrology and Earth System Sciences, 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 Resources Research, 56, e2019WR025 975, https://doi.org/10.1029/2019WR025975, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR025975, 2020.

Kolmogorov, A. N.: Foundations of the theory of probability, Foundations of the theory of probability, Chelsea Publishing Co., Oxford, England, pages: viii, 71, 1933.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, pp. 259–263, https://doi.org/10.1127/0941-2948/2006/0130, publisher: Schweizerbart'sche Verlagsbuchhandlung, 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, Frontiers in Ecology and the Environment, 9, 494–502, https://doi.org/10.1890/100125, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Hydrology and Earth System Sciences, 20, 1069–1079, https://doi.org/10.5194/hess-20-1069-2016, publisher: Copernicus GmbH, 2016.

Messager, 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, Nature Communications, 7, 13 603, https://doi.org/10.1038/ncomms13603, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Hydrology;Limnology Subject_term_id: hydrology;limnology, 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 Resources Research, 53, 8020–8040, https://doi.org/10.1002/2017WR020401, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2017WR020401, 2017.

Mott, R., Vionnet, V., and Grünewald, T.: The Seasonal Snow Cover Dynamics: Review on Wind-Driven Coupling Processes, Frontiers in Earth Science, 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, https://doi.org/10.5067/MODIS/MCD15A3H.006, type: dataset, 2015.

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

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, publisher: Copernicus GmbH, 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, publisher: American Association for the Advancement of Science, 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, Journal of Hydrology, 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, Journal of Glaciology, 60, 537–552, https://doi.org/10.3189/2014JoG13J176, publisher: Cambridge University Press, 2014.

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

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., and Samaniego, L.: Diagnostic Evaluation of Large-Domain Hydrologic Models Calibrated Across the Contiguous United States, Journal of Geophysical Research: Atmospheres, 124, 13 991–14 007, https://doi.org/10.1029/2019JD030767, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Hydrological Processes, 17, 3785–3789, https://doi.org/10.1002/hyp.5167, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.5167, 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, Advances in Water Resources, 22, 367–398, 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, Advances in Water Resources, 23, 15–39, 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, Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, publisher: Copernicus GmbH, 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, Hydrology and Earth System Sciences, 24, 5015–5025, https://doi.org/10.5194/hess-24-5015-2020, publisher: Copernicus GmbH, 2020.

Rutter, A. J., Kershaw, K. A., Robins, P. C., and Morton, A. J.: A predictive model of rainfall interception in forests, 1. Derivation of the model from observations in a plantation of Corsican pine, Agricultural Meteorology, 9, 367–384, 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, Journal of Applied Ecology, 12, 367–380, https://doi.org/10.2307/2401739, publisher: [British Ecological Society, Wiley], 1975.

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

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E. H., Warrach-Sagi, K., and Attinger, S.: Toward seamless hydrologic predictions across spatial scales, Hydrology and Earth System Sciences, 21, 4323–4346, https://doi.org/10.5194/hess-21-4323-2017, publisher: Copernicus GmbH, 2017.

850   Savenije, H. H. G. and Hrachowitz, M.: HESS Opinions <q>Catchments as meta-organisms – a new blueprint for hydrological modelling</q>, Hydrology and Earth System Sciences, 21, 1107–1116, https://doi.org/10.5194/hess-21-1107-2017, publisher: Copernicus GmbH, 2017.

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

855   Schellekens, J., Verseveld, W. v., Visser, M., hcwinsemius, laurenebouaziz, tanjaeuser, sandercdevries, 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, https://doi.org/10.5281/zenodo.4291730, 2020.

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

860   Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H.: Upper and lower benchmarks in hydrological modelling, Hydrological Processes, 32, 1120–1125, https://doi.org/10.1002/hyp.11476, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Hydrology and Earth System Sciences, 19, 4317–4326, https://doi.org/10.5194/hess-19-4317-2015, publisher: Copernicus GmbH, 2015.

865   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, Hydrology and Earth System Sciences, 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.

870   Sorooshian, S. and Gupta, V. K.: Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness, Water Resources Research, 19, 260–268, https://doi.org/10.1029/WR019i001p00260, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens,
875   M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geoscientific Model Development, 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, publisher: Copernicus GmbH, 2018.

Tanaka, T. and Tachikawa, Y.: Testing the applicability of a kinematic wave-based distributed hydrological model in two climatically contrasting catchments, Hydrological Sciences Journal, 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 Resources
880   Research, 41, https://doi.org/10.1029/2005WR003995, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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 Resources Research, 42, https://doi.org/10.1029/2004WR003778, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2004WR003778, 2006.

van Verseveld, W., Visser, M., Bootsma, H., Boisgontier, H., and Bouaziz, L.: Wflow.jl, https://doi.org/10.5281/zenodo.5384924, 2021.

885   Vertessy, R. A. and Elsenbeer, H.: Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization, Water Resources Research, 35, 2173–2187, https://doi.org/10.1029/1999WR900051, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, publisher: Copernicus GmbH, 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 Resources Research, 38, 48–1–48–13, https://doi.org/10.1029/2001WR001118, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2001WR001118, 2002.

Wagener, T. and Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, Journal of Hydrology, 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, Geoscientific Model Development, 14, 3159–3184, https://doi.org/10.5194/gmd-14-3159-2021, publisher: Copernicus GmbH, 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, Scientific Data, 5, 180 118, https://doi.org/10.1038/sdata.2018.118, bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Data publication and archiving;Publication characteristics Subject_term_id: data-publication-and-archiving;publication-characteristics, 2018.

Wood, E. F., Sivapalan, M., Beven, K., and Band, L.: Effects of spatial variability and scale with implications to hydrologic modeling, Journal of Hydrology, 102, 29–47, 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 Resources Research, 47, https://doi.org/10.1029/2010WR010090, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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 Resources Research, 55, 5053–5073, https://doi.org/10.1029/2019WR024873, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR024873, 2019.

Zehe, E., Ehret, U., Blume, T., Kleidon, A., Scherer, U., and Westhoff, M.: A thermodynamic approach to link self-organization, preferential flow and rainfall–runoff behaviour, Hydrology and Earth System Sciences, 17, 4297–4322, https://doi.org/10.5194/hess-17-4297-2013, publisher: Copernicus GmbH, 2013.
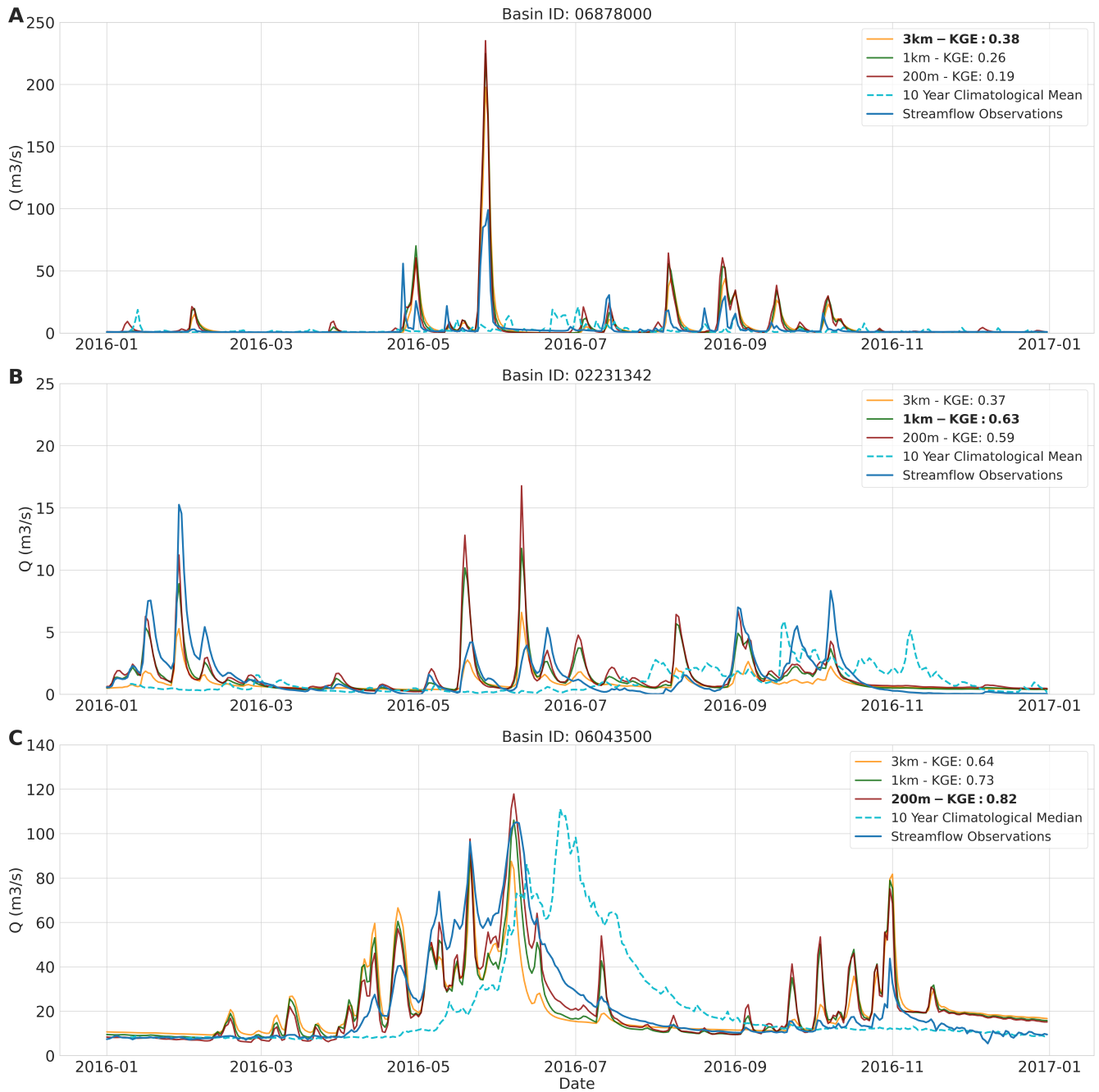
Figure 7. Evaluation period results. 3 example hydrographs showing the last year of the evaluation period. The 3km (orange), 1km (~~green~~red), and 200m (~~red~~green) model instance streamflow estimates at the basin outlet are shown. In blue the streamflow observations and in dotted cyan the 10 year calendar day climatology of the statistical benchmark. (a) Basin ID: 06878000. (b) Basin ID: 02231342. (c) Basin ID: 06043500.
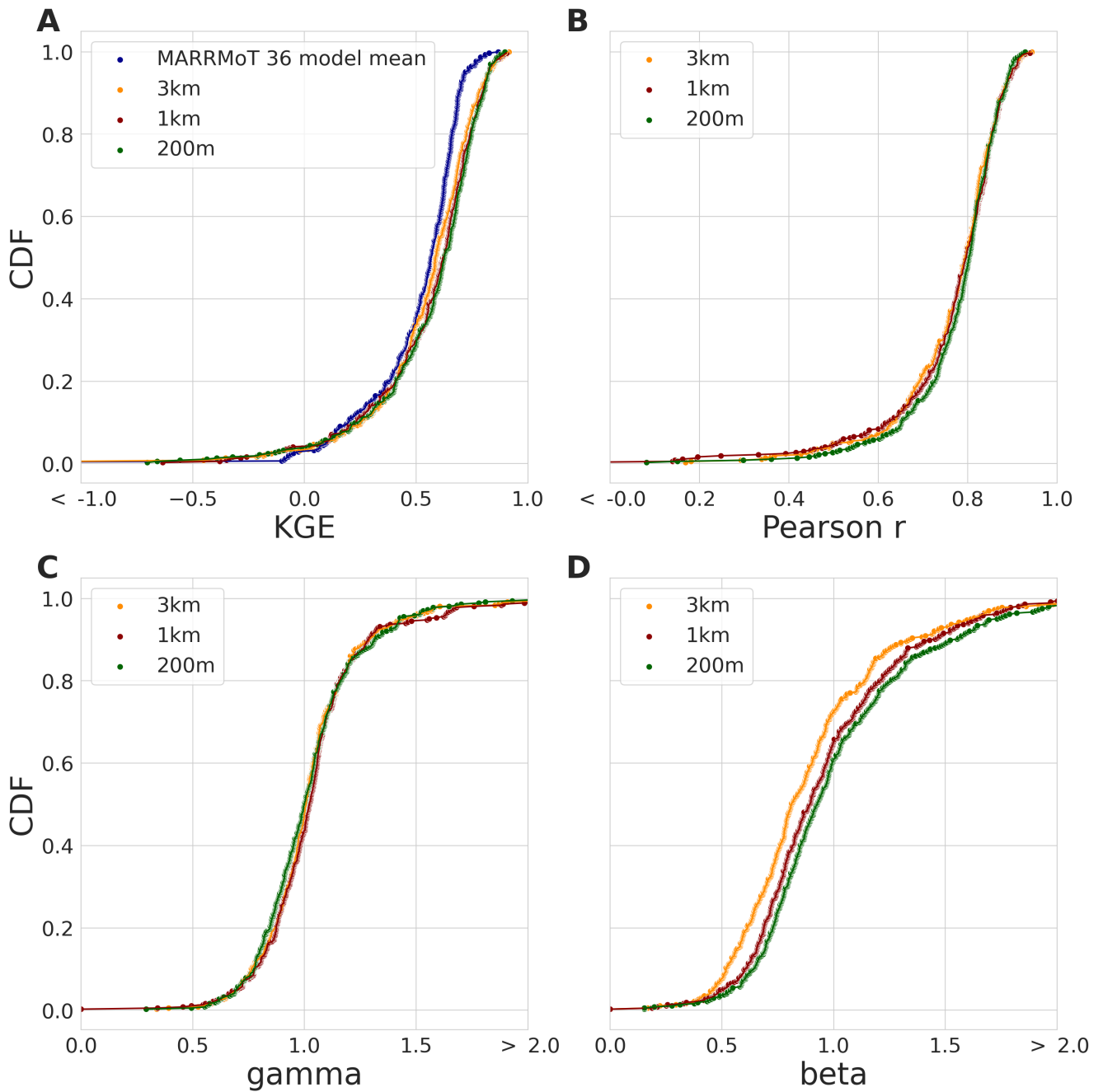
Figure 8. Evauation period results. The CDF based on the modified KGE scores and its 3 individual components for the 454 selected basins. In (a) The modified KGE score with in blue the MARRMoT 36 mean results, in yellow the 3km, in red the 1km, and in green the 200m model instances. (b) The Pearson r correlation component. (c) The gamma variability component. (d) The beta bias component.
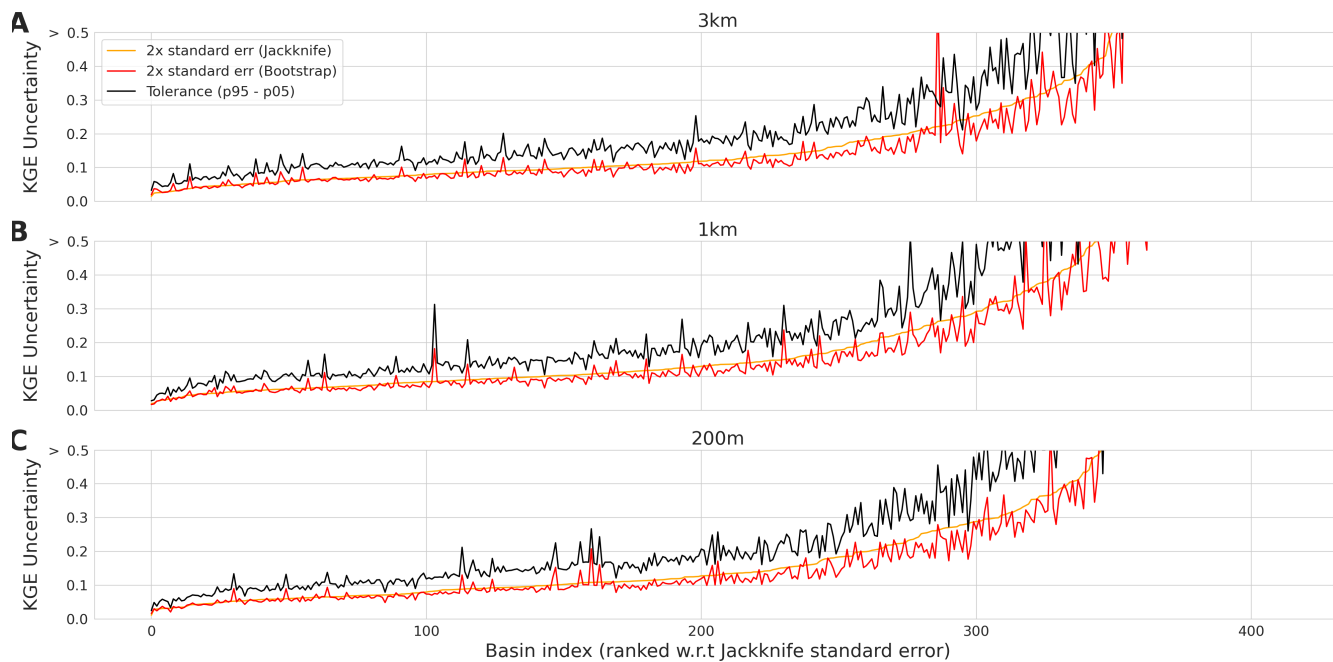
Figure 9. Evaluation period results. The bootstrap and jackknife-after-bootstrap results of the sampling uncertainty of the KGE score. In orange the 2x standard error of the jackknife method, in red the 2x standard error of the bootstrap method, and in black the tolerance interval obtained by subtracting the 5th percentile from the 95th percentile. 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 3km model instance results. b. The 1km model instance results. c. The 200m model instance results.
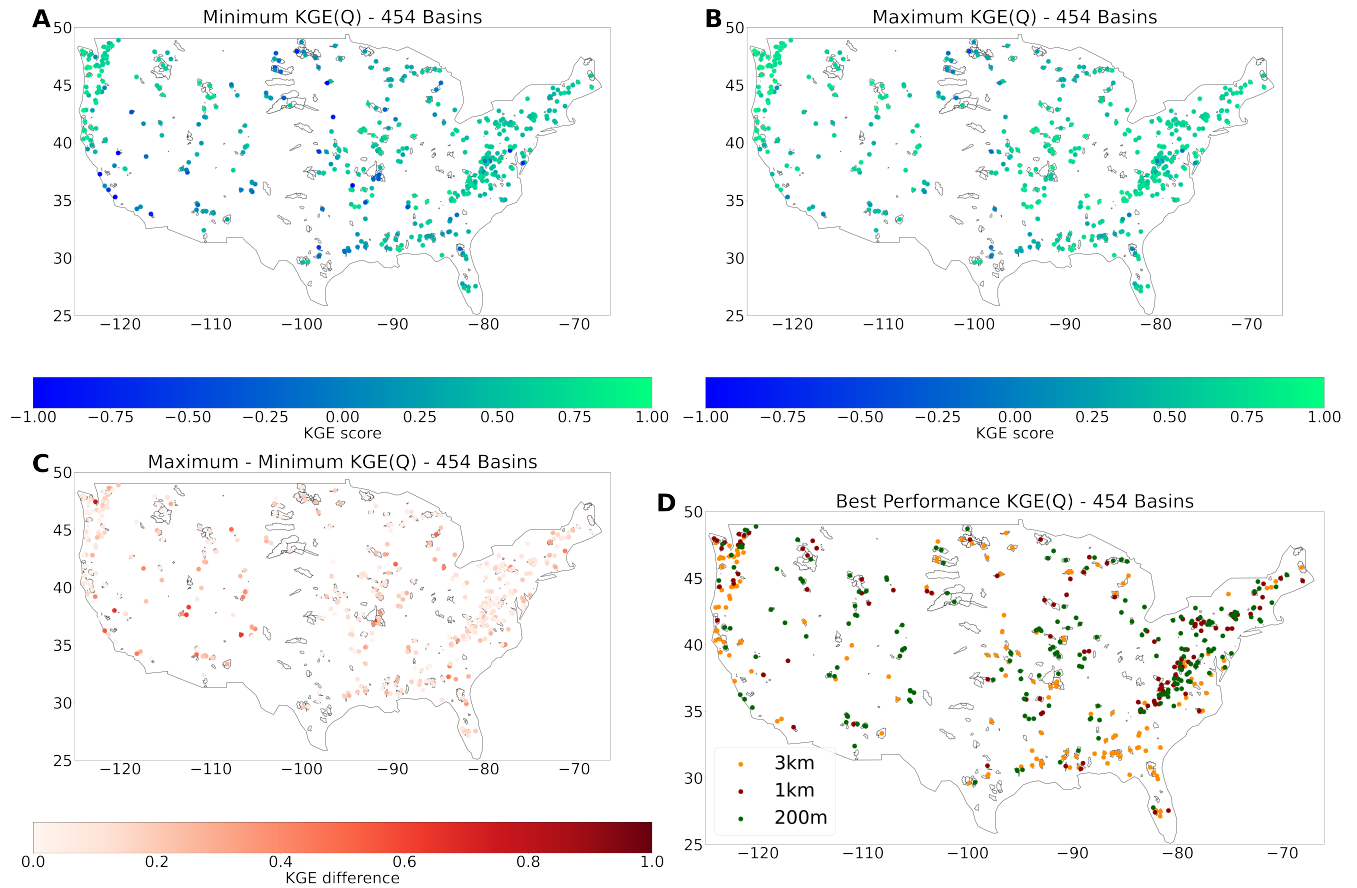
Figure 10. Evaluation period results. (a) Minimum KGE score of the model instances. (b) Maximum KGE score of the model instances. (c) The difference between minimum and maximum KGE scores. (d) Best performing model instance based on KGE score for the evaluation period with 3km in orange, 1km in red, and 200m in green.
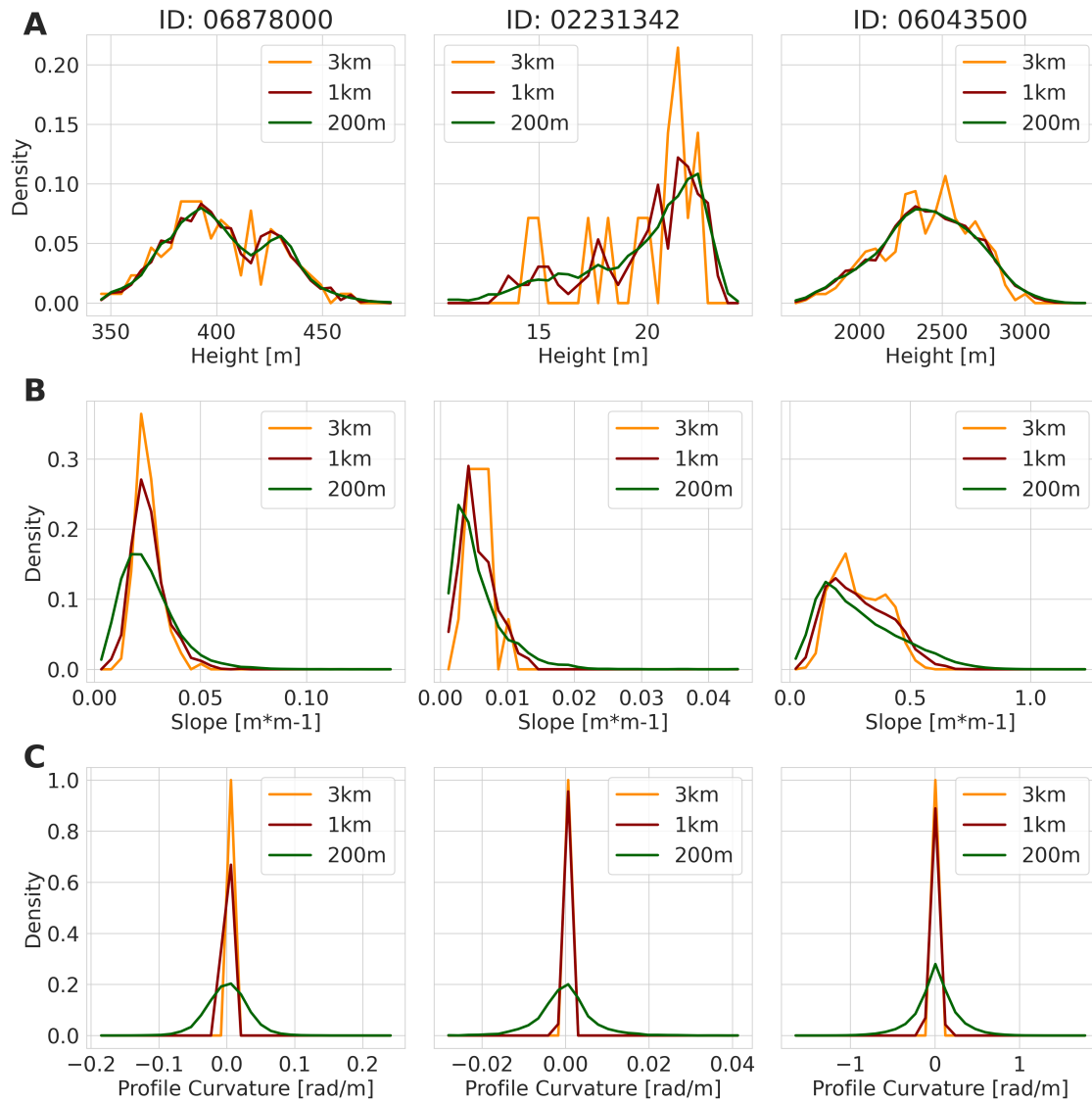
Figure 11. 3 example basins that represent poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). (a) PDF of height distribution for the 3km (orange), 1km (red) and 3km (green) model instances. (b) PDF of the slope distributions of the model instances per basin. (c) 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.
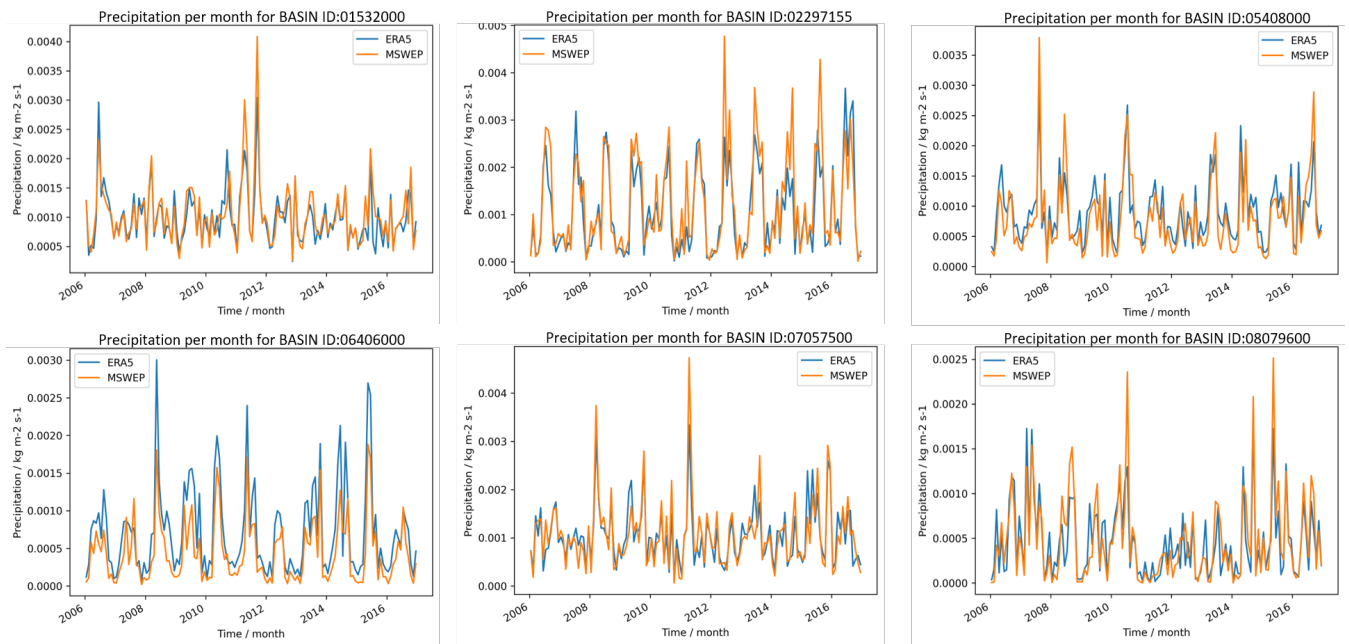
Figure A1. Evaluation period results. ERA5 and MSWEP forcing comparison for 6 basin in the CAMELS Dataset. Monthly precipitation values for the evaluation period are shown with in blue ERA5 and in orange MSWEP.
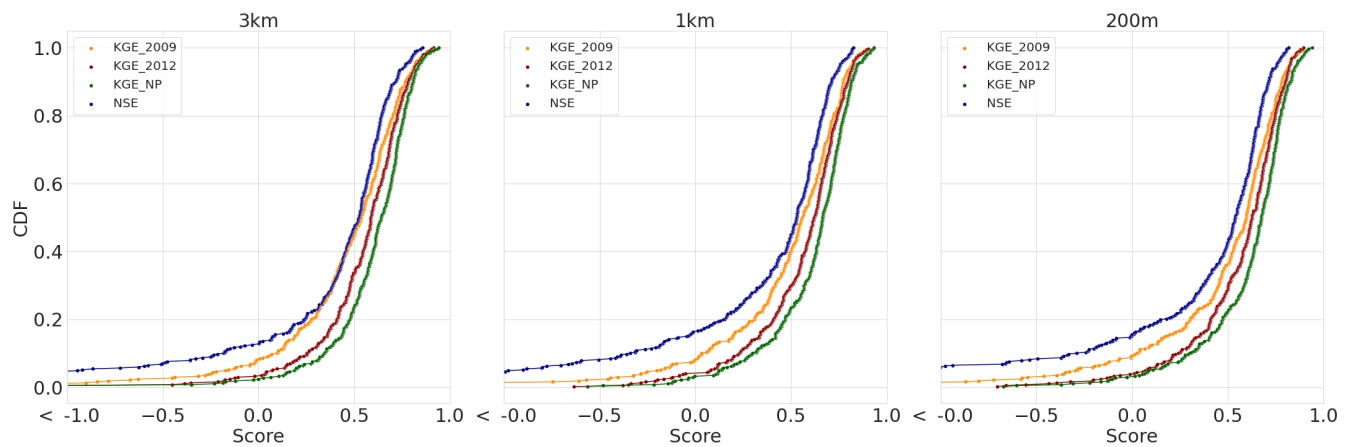
Figure A2. ~~CDFS~~ Evaluation period results. CDFs of multiple objective functions for the 3 model instances. With in ~~blue~~ orange KGE 2009, in ~~orange~~ red KGE 2012, in ~~red~~ green KGE NP, and in ~~green~~ blue NSE.