https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

Hydrology and Earth System Sciences
Open Access
Discussions

Large-sample assessment of spatial scaling effects 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 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 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 only a statistical difference between the 3km and 200m streamflow estimates. However, results indicate strong locality in scaling behaviour 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. 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

In recent years, the hydrologic modelling community is making an effort to 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 (Wood et al., 2011; Beven and Cloke, 2012; Bierkens et al., 2015). These challenges include scaling issues such as; (1) 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 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 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 complicating the issues are processes that might change in level of complexity at aggregated (time and spatial) scales due to self-organizing principles of the ecosystem (e.g. Bak, 1996; Zehe et al., 2013; Savenije and Hrachowitz, 2017).

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 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 resolutions.

Multiple studies have tested 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. 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 resolution. This study made it apparent that a large sample of hydrological diverse basins should be considered when investigating spatial scaling effects. To our knowledge there are no studies that have looked into scaling effects 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 for the horizontal conductivity fraction (KsatHorFrac) is yet available, 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 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 suggested causes of the inconsistent streamflow across scales as shown by Imhoff et al. (2020)).

In this study we quantify the effects of spatial scaling 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.

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

2 Methodology

2.1 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.

Of the 671 basins we ran each of the 3 model instances (i.e. 3km, 1 km and 200m resolution) successfully for 567 basins. Failed runs were either caused by errors during parameter derivation, errors during run time, or missing streamflow observations. If a single model instance failed the basin was excluded from further analysis. Figure 1 shows the locations of the successful and failed model runs.

2.1.2 Streamflow Observations

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. 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_notebooks).

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 ($\tilde{1}$1km) and 3-hourly temporal resolution for the period 1979 to 2017. The temperature variable was obtained from the ERA5 reanalyses data set (Hersbach et al., 2020). The data is available at 0.25 degrees ($\tilde{3}$1km) 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 of ERA5 did not produce desirable streamflow estimates. Results of this analysis are included in the data repository. Switching to the MSWEP precipitation product improved streamflow estimates throughout the case study area.

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 val-

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

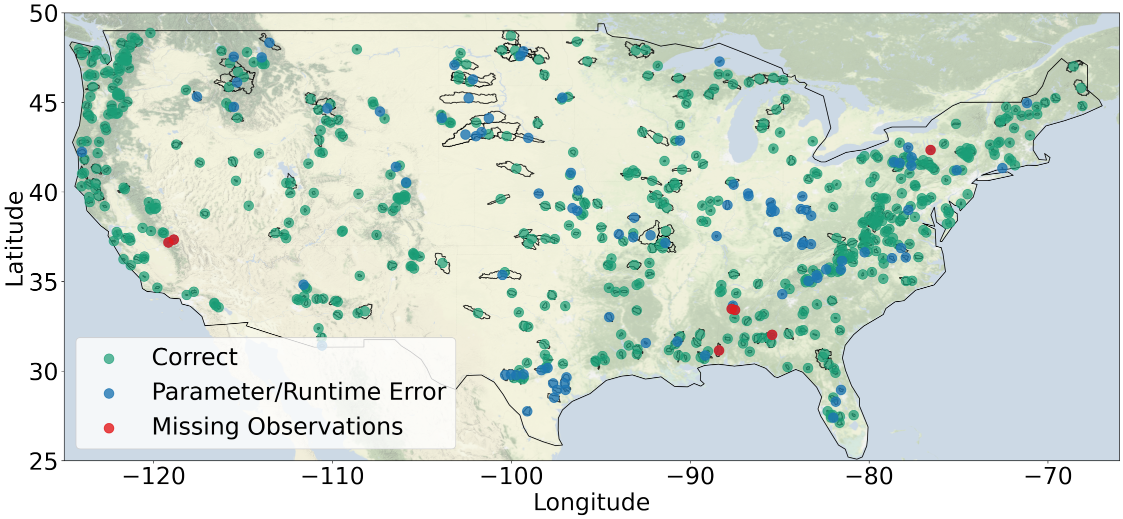

Figure 1. Basin locations of the CAMELS data set with in green the successful model runs. In blue failed runs due to parameter or run time errors and in red due to missing streamflow observations. Basemap made with Natural Earth.

ues. 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 are disaggregated using the (bi)linear method and subsequently used to calculate potential evapotranspiration. The code for all these pre-processing steps is included in the Jupyter Notebooks made available with this manuscript (DOI:10.5281/zenodo.5724512).

2.2 Model Experiment Setup

2.2.1 The wflow_sbm model (v.2020.1.2)

Wflow_sbm is available as part of the wflow 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 fluxes that are part of the

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

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. 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 physical 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 (< 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. The main differences of wflow_sbm with Topog_SBM are:

- 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).
- 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

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.

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)

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

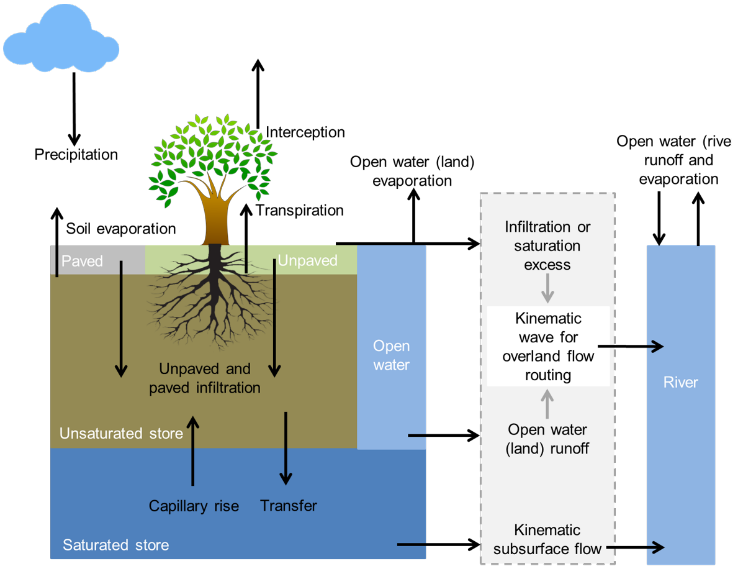
Precipitation
Interception
Transpiration
Soil evaporation
Open water (land) evaporation
Open water (river) runoff and evaporation
Paved
Unpaved
Unpaved and paved infiltration
Unsaturated store
Open water
Infiltration or saturation excess
Kinematic wave for overland flow routing
Open water (land) runoff
River
Capillary rise
Transfer
Saturated store
Kinematic subsurface flow

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

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

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). 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 this parameter leads to an increased base flow component and reduces peak flow and flashiness. We calibrated the models to match model setups of those used by the users of the hydrological model.

2.2.4 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.

2.3 Analyses of results

2.3.1 Benchmark Selection

To select basins where model performance is 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 resolutions, has a lower KGE than either the median or mean flow, it is considered not suited for this study and removed from the list of basins.

2.3.2 Comparison of 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.

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

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

between the KGE score distributions of the model instances are statistically relevant. Allowing the acceptance or rejection of the hypothesis stating that the differences in streamflow estimates at various spatial resolutions will be small.

3 Results

The results in this section are based on the KGE objective function at the basin outlet for the evaluation period. Results based on the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) 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

To illustrate how model calibration effects 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. 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 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 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 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 4a shows the probability density function (PDF) of the height distributions of the model instances for each of the 3 basins, Figure 4b shows the slope distribution, and Figure 4c the profile curvature distributions.

The height distribution of the DEM in Figure 4a 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 4bc. As can be expected from the height distributions, the slope of the 200m instance has more gentle and steep sloping topography

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

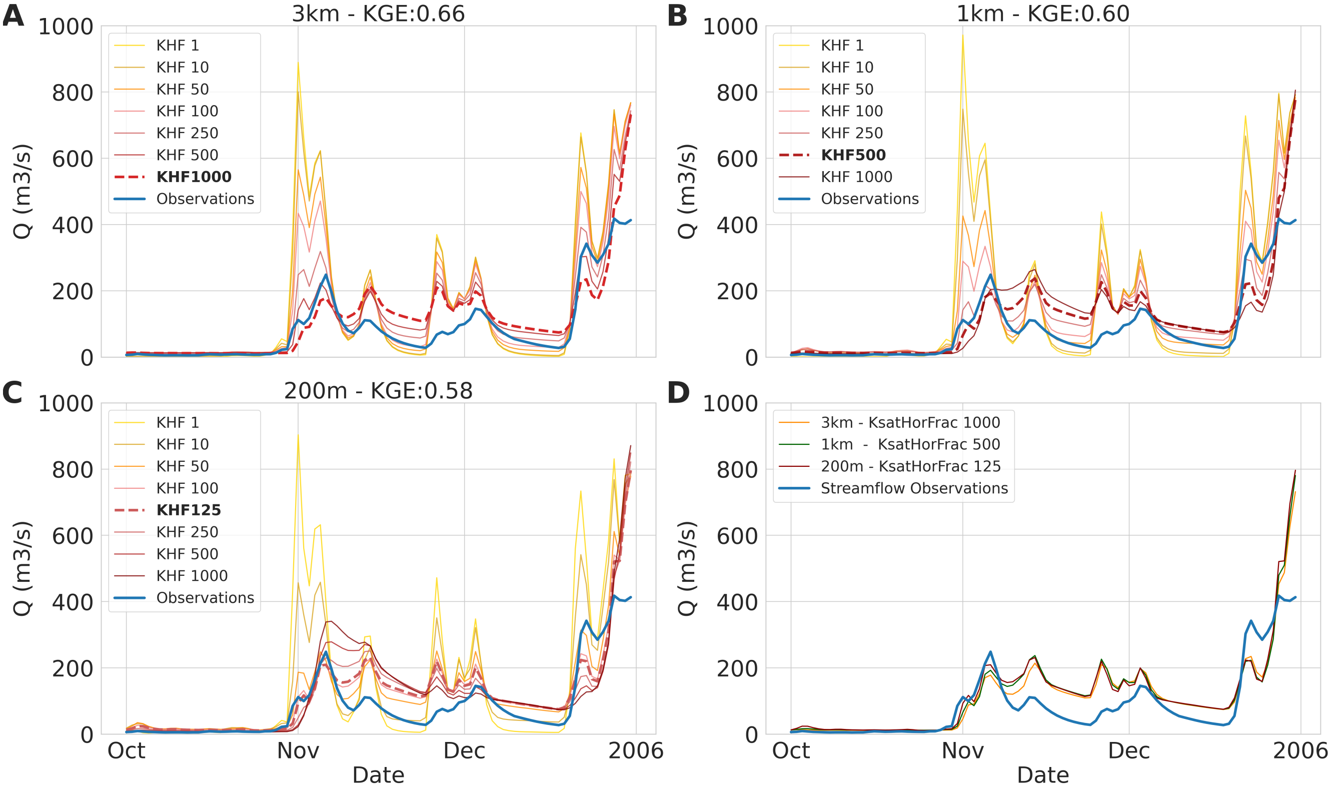

Figure 3. The calibration interval of the KsatHorFrac parameter (KHF) for the 3 model instances at the basin outlet (ID:14301000). (a) 3km model instance. (b) 1km model instance. (c) 200m model instance. Values range from 1 to 1000 (yellow to red), only a sub-selection is shown. Best performing calibration 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.

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 4c 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 2 show small differences between the model instances for each of the 3 basins.

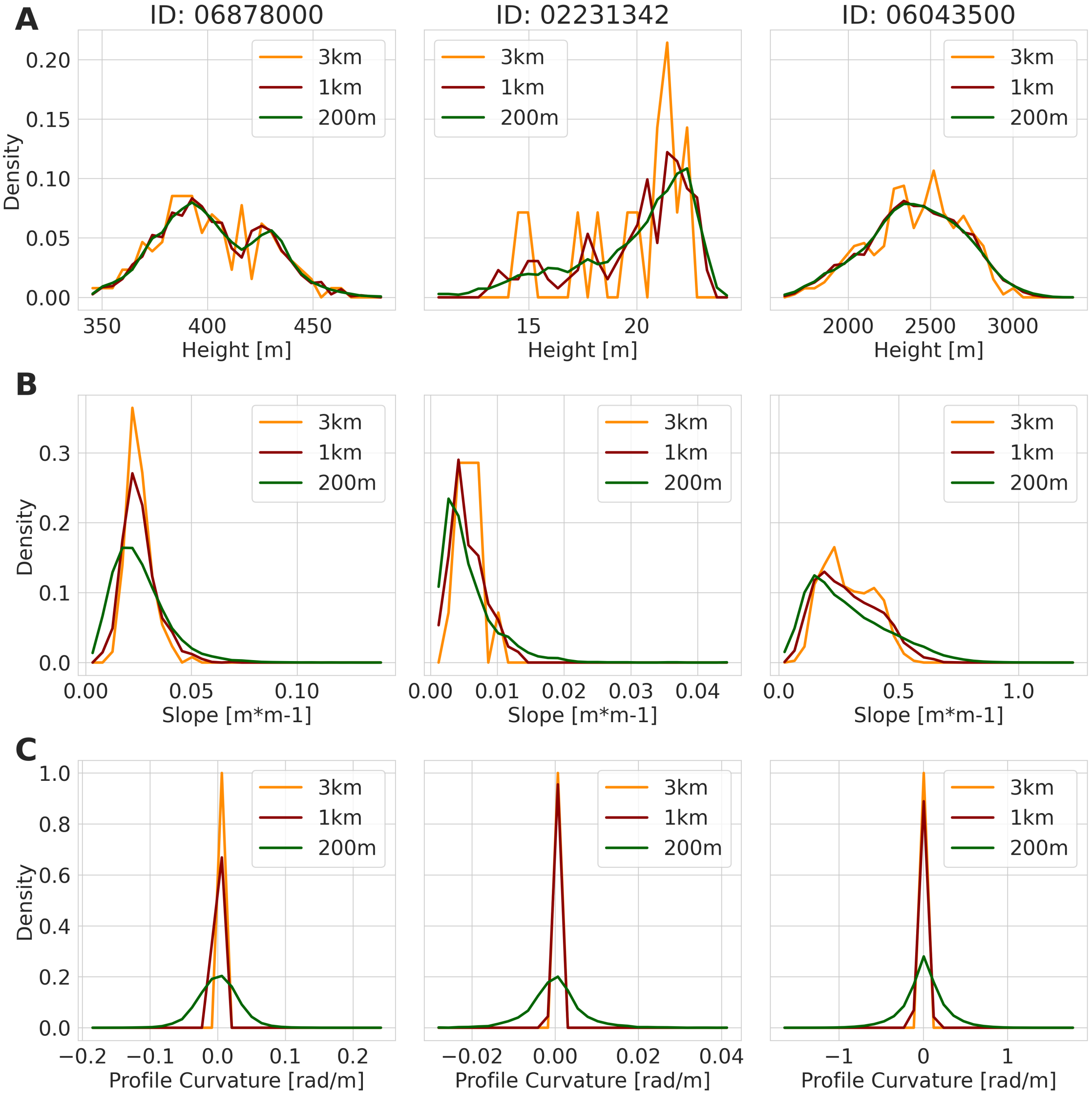

Figure 4. 3 example basins that represent poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). (a) PDF of height distribution for the 3km (orange), 1km (red) and 3km (green) model instances. (b) PDF of the slope distributions of the model 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.

3.3 The effect of spatial scale on streamflow estimates

The same 3 example basins are used to illustrate the differences in streamflow estimates between model instances for the evaluation period. Only the last year of the evaluation period is shown in Figure 5.

Table 2. 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

In the case of poor performance, Figure 5a, 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 5b 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 5c 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.

3.4 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 7a. Figure 7b 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 7c and the distribution of the benchmark KGE scores is shown in Figure 7d. 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).

3.5 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). The KGE score distribution of the mean of 36 hydrological models from 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.

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

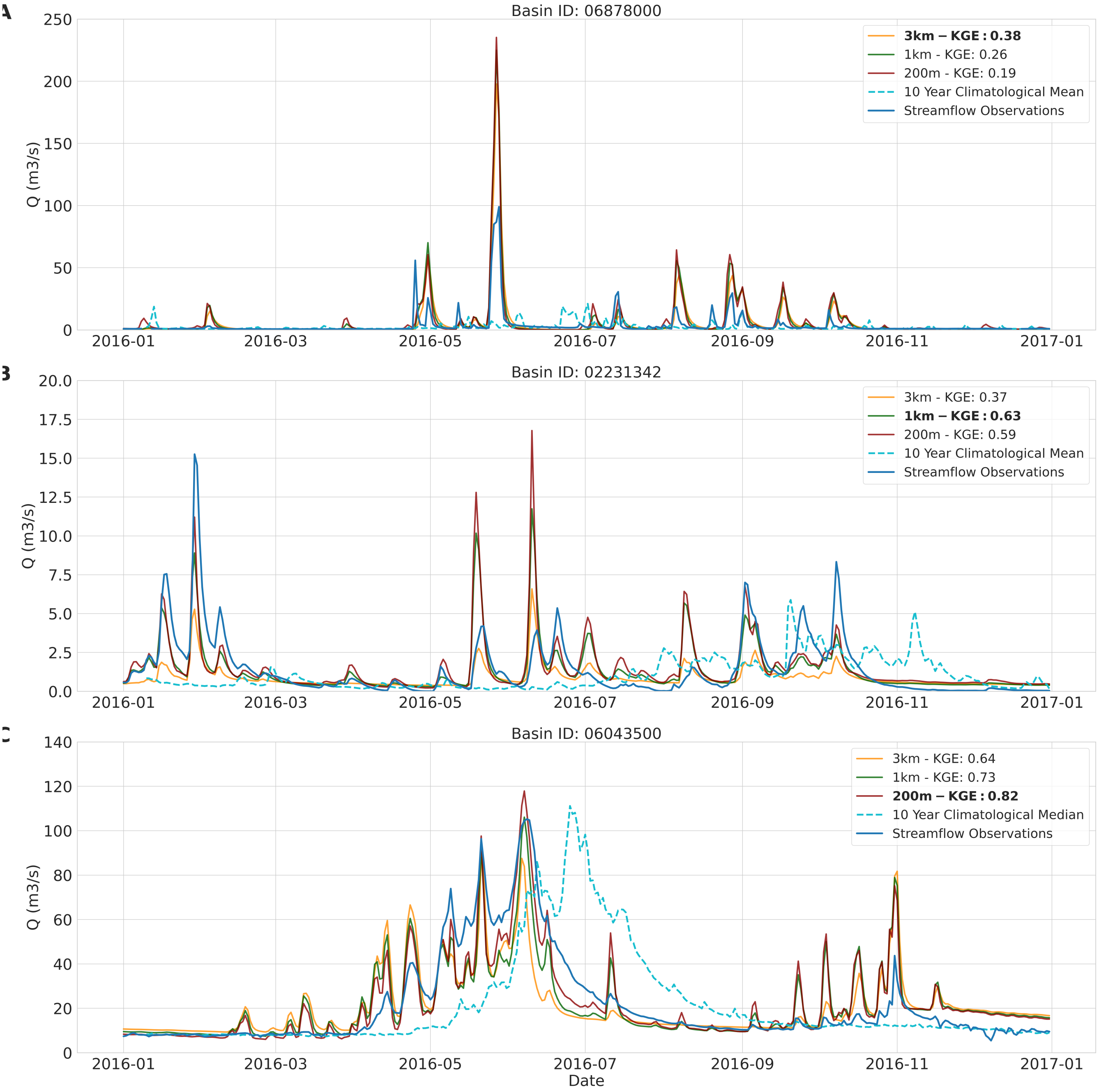

Figure 5. 3 example hydrographs showing the last year of the evaluation period. The 3km (orange), 1km (green), and 200m (red) model instance streamflow estimates at the basin outlet are shown. In blue the streamflow observations and in dotted cyan the 10 year calendar day climatology of the statistical benchmark. (a) Basin ID: 06878000. (b) Basin ID: 02231342. (c) Basin ID: 06043500.

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

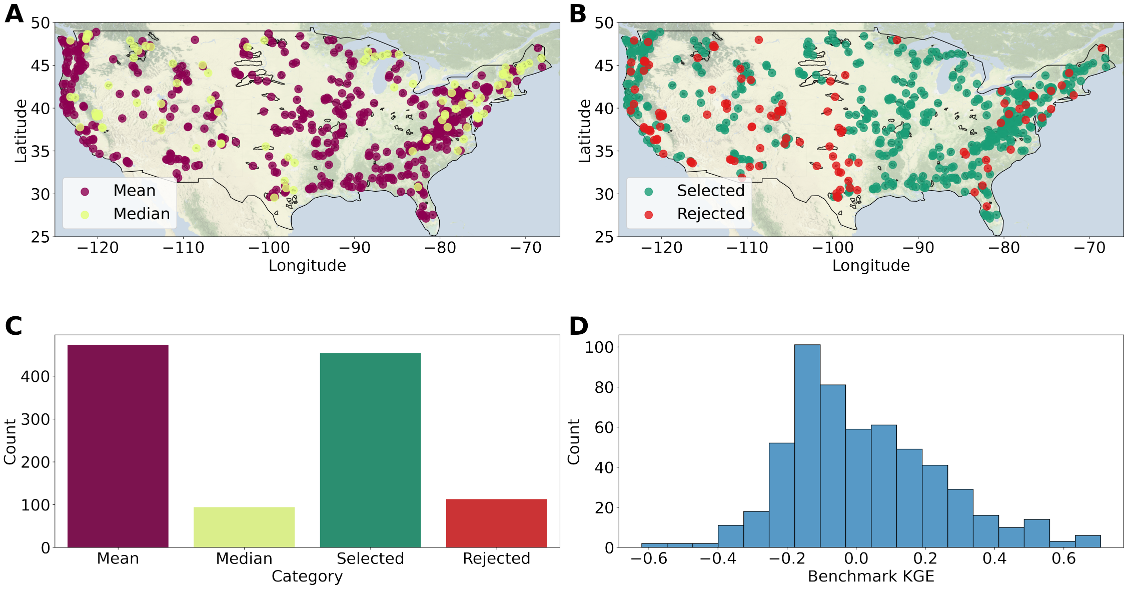

Figure 6. (a) The best performing type of 10 year calendar day climatology, either mean (purple) or median (yellow). (b) The spatial distribution of accepted (green) and rejected (red) basins based on the benchmark. (c) Overview of the amount of basins that are accepted or rejected and the best performing benchmark type. (d) The KGE score distribution of the best performing benchmark type. Basemaps made with Natural Earth.

The mean KGE score distribution of the MARRMoT models (Figure 7, blue) of 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.

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 3 show that the difference between 3km and 200m model instances is statistical relevant at a p-value of 0.01. On a large-sample, this means that increasing the spatial model

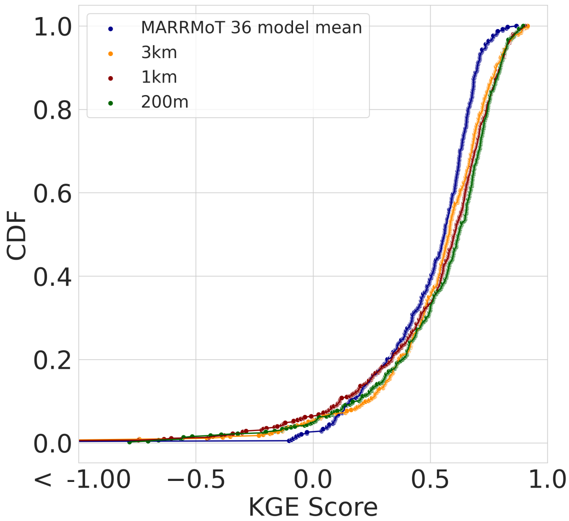

Figure 7. The CDF based on KGE scores for the 454 selected basins. In blue the MARRMoT mean results, in yellow the 3km, in red the 1km, and in green the 200m model instances.

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

CDFs	Kolmogorov-Smirnov Statistic	p-value
3 km - 1 km	0.07	0.21
1 km - 200 m	0.06	0.31
3 km - 200 m	0.10	0.01

resolution from 3km to 1km or 1km to 200m does not lead to significant differences in streamflow performance. However, when changing resolution from 3km to 200m, the distribution of KGE scores is significantly different ($p<0.05$).

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 8 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 8a, 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 8b are located in the Northwest, Rocky Mountains, and Appalachia. These regions are characterised as steep sloping headwater basins. Figure 8c 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. Figure 8d 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

https://doi.org/10.5194/hess-2021-605
Preprint. Discussion started: 1 December 2021

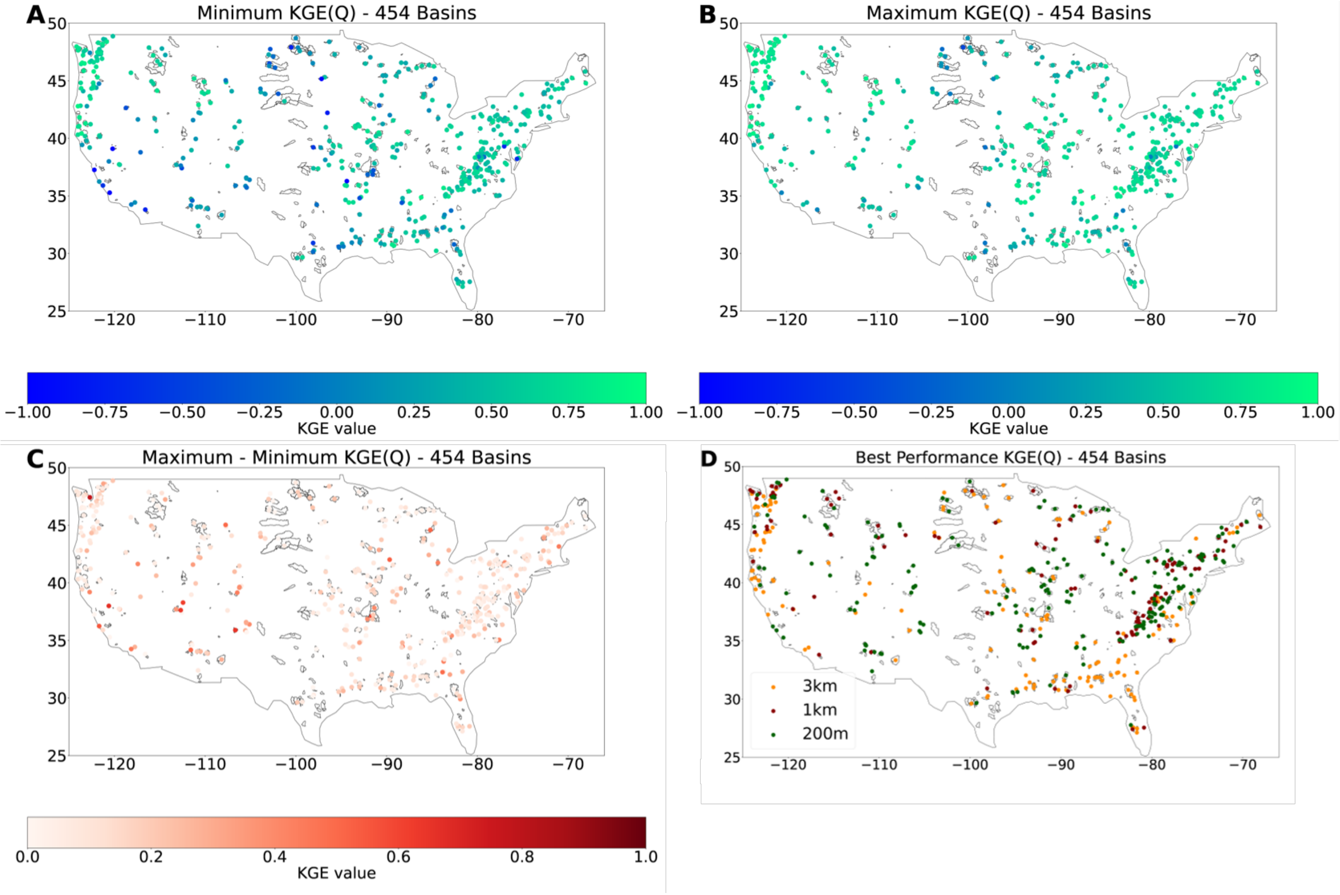
A
Minimum KGE(Q) - 454 Basins
B
Maximum KGE(Q) - 454 Basins
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
KGE value
C
Maximum - Minimum KGE(Q) - 454 Basins
0.0
0.2
0.4
0.6
0.8
1.0
KGE value
D
Best Performance KGE(Q) - 454 Basins
3km
1km
200m

Figure 8. (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.

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.

4 Discussion

We applied an initial benchmark based on streamflow observations for basin selection to identify basins where streamflow estimates are deemed adequate for further analyses. This does not imply that excluded basins are less relevant. Instead it implies that an in-depth model assessment is required to understand why the model is not able to simulate adequate streamflow estimates in these basins. In addition to the benchmark, we added a layer of context by including results from the study by

Knoben et al. (2020). This is an imperfect comparison due to differences in 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 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.

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. We further propose the use of a model experiment environment, such as eWaterCycle (Hut et al., 2021), to generate model results. This allows for similar pre-processing of inputs, standardization of outputs, and reproducible modelling studies. An example of how to apply these steps using the eWaterCycle platform is provided in 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 suitable for publication.

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 7) show that this is the case for the KGE score distributions based on streamflow estimates at the basin outlet. The crossing of the distribution lines is a strong indication that their is disagreement on KGE scores between model instances and that there is no single instance outperforming the rest consistently. For some basins the coarser 3km instance outperformed the finer 200m instance, showing that finer resolution does not automatically lead to improved streamflow estimates.

Next, we applied the KS-statistic to test if there is a significant difference between the distributions of model instances. This is only the case for the model 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 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.

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 are in effect throughout the spatial domain. This is depicted by the differences between the KGE scores of the model instances (Figure 8c). On average these is a 0.22 KGE score difference with extremes of more than 0.5 KGE score difference at multiple basins. 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 8c). 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 instances (Figure 8d) there are only small geographical trends of best model performance in the South and Appalachia.

We conducted a terrain analyses (Figure 4) 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 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. 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 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 results in a broader distribution of slopes that effects the volume and timing of streamflow estimates. An in-depth assessment of the 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 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). It is of interest to investigate this effect on streamflow estimates and to determine the role of native spatial meteorological forcing resolution.

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.

In this study we did not investigate individual basins to avoid biased selection of case study areas. However, we suggest that future work investigates the basins that show large differences in lateral fluxes and poor model performance. By including multiple evaluation data products (e.g, soil moisture, evaporation, gravitational anomaly) conclusions can be made on whether increasing spatial resolution leads to increased model fidelity. 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 on the streamflow estimates of the distributed wflow_sbm hydrological model. Distributions of model instance KGE score results were tested for significant differences. 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.
- Results show large differences in maximum and minimum KGE scores with an average of 0.22 between model instances throughout the CONUS.
- There is no single best performing model resolution across the domain.
- 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. 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

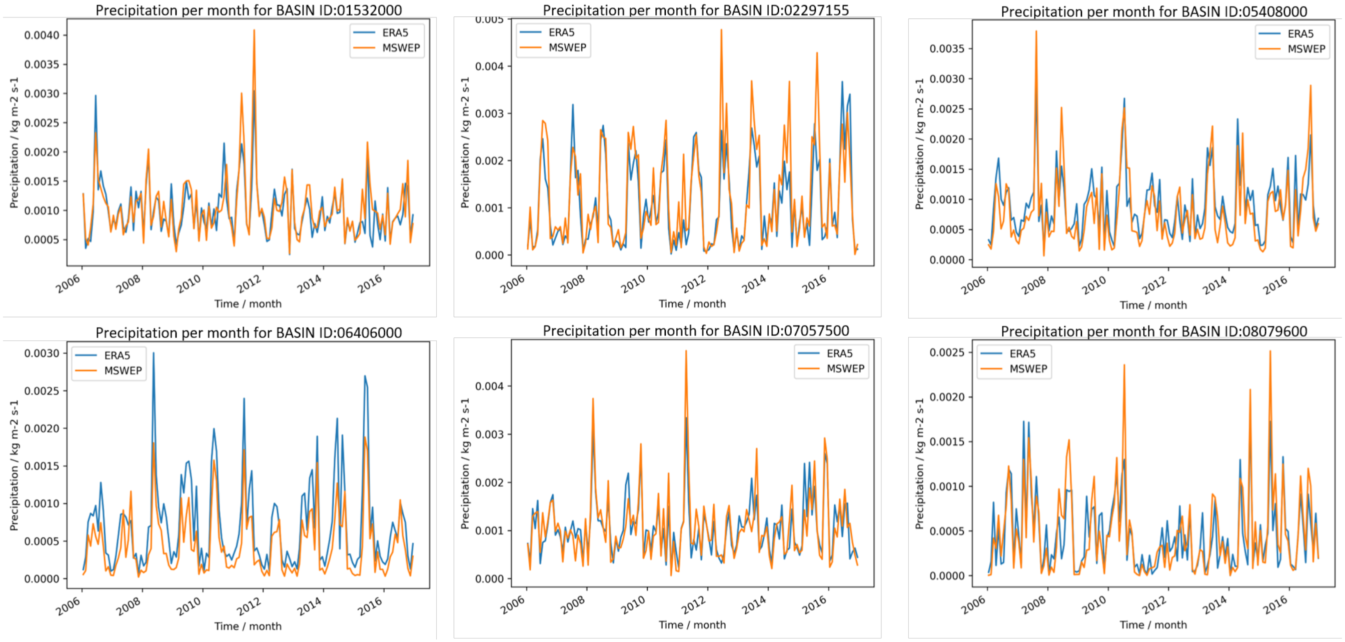

Figure A1. ERA5 and MSWEP forcing comparison for 6 basin in the CAMELS Dataset. Monthly precipitation values for the evaluation period are shown with in blue ERA5 and in orange MSWEP.

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

BASIN ID	Resolution	ERA5 - KGE 2012	MSWEP - KGE 2012	ERA5 - NSE	MSWEP - 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

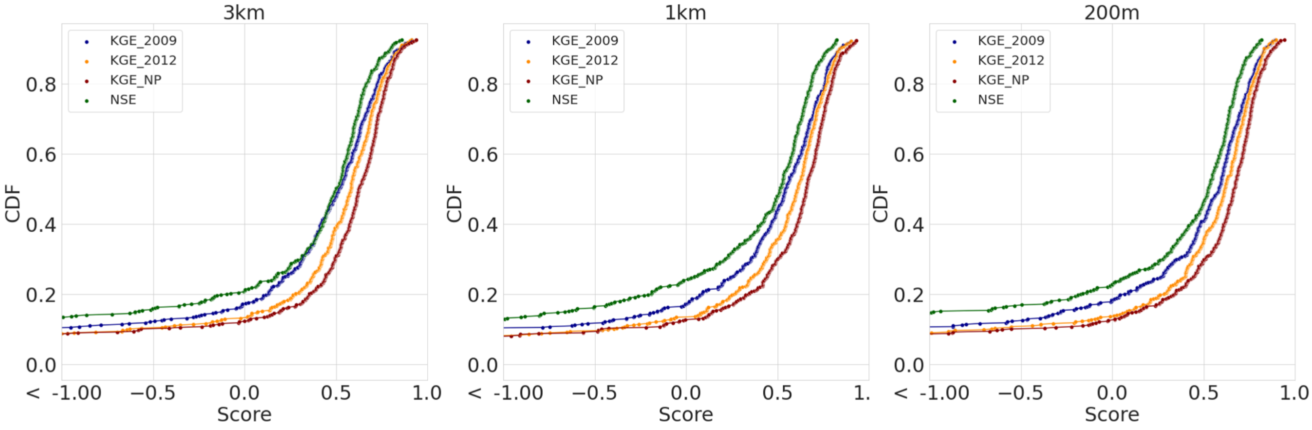

Figure A2. CDFS of multiple objective functions for the 3 model instances. With in blue KGE 2009, in orange KGE 2012, in red KGE NP, and in green NSE.

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. 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.

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., 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.

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.

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.

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 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.

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: Elsevier, 2009.

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.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, 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.

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., 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.

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.

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.

Savenije, H. H. G. and Hrachowitz, M.: HESS Opinions <q>Catchments as meta-organisms &ndash; 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.

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.

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.

Smirnov, N.V.: Estimate of Deviation between Empirical Distribution Functions in Two Independent Samples., Bulletin Moscow University, 2, 3-16., 1933.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5&thinsp;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 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.

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.

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., 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&ndash;runoff behaviour, Hydrology and Earth System Sciences, 17, 4297–4322, https://doi.org/10.5194/hess-17-4297-2013, publisher: Copernicus GmbH, 2013.