Effect of three pillars on hydrological model calibration: data length, spin-up period and spatial model resolution

Ömer Ekmekcioğlu [1*], Mehmet Cüneyd Demirel [1], Martijn J. Booij [2]

[1] Department of Civil Engineering, Istanbul Technical University, 34469 Maslak, Istanbul, Turkey.

[2] Water Engineering and Management, Faculty of Engineering Technology, University of Twente, Enschede 7500 AE, The Netherlands

Corresponding author: Ömer Ekmekcioğlu (omer.ekmekcioglu@itu.edu.tr)

Key Points:

- A systematic approach is presented to identify appropriate calibration data length, spin-up period and spatial model resolution
- Dependency of model performance on data length, spin-up period and spatial resolution of the model schematization is revealed for the Moselle River
- Three user-defined pillars in modelling should not be overlooked due to trade-off between computational costs and model performance

Abstract

In general, calibration of a hydrologic model is essential to better simulate the basin processes and behaviour by fitting the model simulated fluxes to observed fluxes. A major challenge in the calibration process is to choose the appropriate length of the observed data series and spatio-temporal resolution of the model schematization. We present a multi-case calibration approach for determining three pillars of an optimum hydrological model configuration: calibration data length, spin-up period and spatial resolution of the hydrological model. The approach is evaluated for the Moselle River basin using calibration and validation results from the spatially distributed meso-scale Hydrological Model (mHM) for 105 different cases representing the combinations of three calibration data lengths, seven spin-up periods and five spatial model resolutions. A metaheuristic global optimization method, i.e. Dynamically Dimensioned Search (DDS) algorithm, and a well-known hydrological performance metric, i.e. Nash Sutcliffe Efficiency (NSE), are utilized for each of the 105 calibration cases. The results show that a 10-year calibration data length, 2-year spin-up period and a 4 km model resolution are appropriate for the Moselle basin to reduce the computational burden. Analyzing the combined effects further allowed us to understand the interactions of these three usually overlooked pillars in hydrological modeling.

1 Introduction

Hydrological models are crucial tools to evaluate physical processes and quantify water balance components in a catchment. They can be classified according to the amount of physics incorporated as empirical (or data-driven), conceptual and physically-based models. The focus in this study is on physically-based regarding the amount of physics and fully-distributed regarding the spatial resolution of the models. Obviously, the choice of the model type together with data availability such as the spatial resolution of inputs, the length of the spin-up period and the parameter calibration strategy all affect the model performance (Blöschl & Sivapalan, 1995). The determination of all these aspects in a calibration framework is related to appropriate modeling in hydrology and should be based on the modeling objective, data availability and a systematic analysis of the model-catchment interaction (Booij, 2005). We focus on user-defined options in hydrological modelling as we are interested in identifying the appropriate calibration data length, spin-up period and spatial model resolution in the Moselle River basin.

The calibration process, which has utmost importance to minimize the parameter uncertainty (Sreedevi & Eldho, 2019; Westerberg et al., 2020), is described as the optimization of uncertain parameter values in the model to obtain sufficient accuracy in model outcomes (Simunek et al., 2012). Since calibration can be performed by trial-and-error for different conditions, i.e. manual calibration (Gelleszun et al., 2017), and also with mathematical algorithms, i.e. automatic calibration (Madsen, 2003), time-efficiency is a major challenge. The main constraint in determining the calibration period is the availability of data, i.e. long time-series of runoff or other model output or state variables (Sorooshian et al., 1997). In general, using 20-year data for the calibration period is assumed to be sufficient for large basins to account for climatological and hydrological variability (Epstein et al., 1998). Although data records for large basins might be available for more than 30 years, keeping the calibration period as long as possible is computationally inefficient and not always meaningful, in particular when climatic or other trends are present in the time series and the model only should be calibrated on the most representative (i.e. most recent) time period (Daggupati et al., 2015). For instance, Perrin et al. (2007) found that a much smaller number of random days (~300 days) is sufficient for calibration of models with a small number of parameters.

In different studies, even data periods of 10 years or less have been used considering both computational resources and limited data availability (Andersen et al., 2001; Kim et al., 2018). Zheng et al. (2018) analyzed the impact of different calibration periods on model results using data-driven techniques. They concluded that the model performance may increase by considering temporal variability and extreme events in the calibration process. In addition, a number of studies has confirmed that quality of data increases calibration performance in distributed hydrological models (Beck et al., 2017; Herman et al., 2018; Näschen et al., 2018). Raihan et al. (2020) evaluated the calibration performance of hydrological models according to different performance criteria and showed that the simulations were not considerably successful particularly for extreme low flows due to the limited temporal variability and poor data quality of the calibration data.

Another factor affecting the calibration performance of hydrological models is the length of the spin-up period, which provides the required initial model state (Yang et al., 1995). The required spin-up period highly depends on the input data of the catchment and the hydrological response (Rodell et al., 2005). In addition, determining the optimum spin-up period is essential, since both shorter and longer spin-up periods may have negative effects on the calibration performance. A shorter spin-up period inevitably leads to a low (even wrong) performance evaluation, whereas a longer spin-up period can lead to a waste of the data and misinterpretation of the results (Ajami et al., 2014). Practitioners generally consider the first two or three years as acceptable as spin-up period depending on the model structure. There have been studies using only a spin-up period of one year for lumped models (Rahman et al., 2016), semi-distributed models (Abdo et al., 2009; Xu et al., 2013) and distributed models (Cuo et al., 2006; Lohmann et al., 1998; Revilla-Romero et al., 2016). Although there is common sense that the spin-up period varies from one year to several years up to ten years (Shi et al., 2008), no consensus has been reached in this regard (Kim et al., 2018). Sood et al. (2013) performed simulations with a monthly time step, since they had monthly streamflow observations, and the first two years of a 13-year data period have been used as spin-up period, while the remaining 11 years have been utilized for model calibration. Ashraf (2013) performed simulations on a monthly basis as well and divided the entire data set into two periods with six years as spin-up period and ten years as calibration period. With a few exceptions, studies conducted to identify the optimum spin-up period surprisingly did not attract the research community's attention, particularly for physically-based distributed hydrological models.

Besides, heterogeneous land surface conditions require a sufficiently long spin-up period (Shrestha & Houser, 2010). Ajami et al. (2014) emphasize the importance of a multi-criteria approach, which includes the groundwater storage, unsaturated zone storage, depth to water table, root zone storage, discharge, snow water equivalent and energy fluxes, in determining the spin-up period of integrated hydrological models. The length of spin-up periods also depends on the initial soil moisture content, soil depth, soil and vegetation type and groundwater storage at the start of the simulations, in addition to the temperature and rainfall forcings (Cosgrove et al., 2003). With a method based on relative changes in monthly groundwater storages, Ajami et al. (H. Ajami et al., 2014) presented a hybrid approach on the basis of integration of ParFlow, which is an integrated hydrological model, and the empirical depth-to-water-table function, to satisfy state equilibrium conditions. They reduced the spin-up period by approximately 50% (from 20 years to 10-12 years) compared to the conventional continuous recursive simulation approach, which is widely employed for the determination of spin-up periods in land surface models.

Regardless of the model complexity, another issue which has a significant impact on hydrological model performance is the spatial model resolution (Koren et al., 1999). The spatial resolution to be used in a model is not only related to the availability of meteorological input data but also to the computational resources (Sood & Smakhtin, 2015). Accordingly, simulation performance may either increase or decrease depending on the spatial resolution (Booij, 2002; 2005; Bucchignani et al., 2016; Pang et al., 2020). However, in some cases, a considerable change is not observed indicating that the model structure is suitable for all resolutions (Merz et al., 2009). In addition, the spatial variability of storm events also has an influence on the appropriate spatial resolution of the model. Lumped models may perform accurately with a spatially uniform input distribution, while they may need a higher spatial resolution (e.g. sub-basins) in the case of a non-uniform spatial input distribution (Tian et al., 2020). Pang et al. (Pang et al., 2020) evaluated the precipitation model input, both temporally and spatially, based on the differences of various open access precipitation products.. In semi-distributed conceptual models, the spatial resolution is determined based on the sub-basin distribution. Distributed models provide distributed outputs since spatial heterogeneity is taken into account (Dehotin & Braud, 2008). Etchevers et al. (2001) performed simulations for spatial resolutions of 1 km, 8 km and 46 km using the soil-vegetation-atmosphere transfer (SVAT) model. They obtained mediocre simulation results for the 46 km resolution, whereas flash-flood events were better captured in the model with a 8 km resolution. Chen et al. (2017) employed the Liuxihe model, i.e. a physically based distributed hydrological model, to investigate flood events in Liujiang River basin, China, which covers an area of about 60000 km^2. They calibrated the model using Particle Swarm Optimization (PSO) for a total of 29 flood events. Considering five different spatial model resolutions, i.e. 200, 400, 500, 600 and 1000 m, they concluded that the results for the 1 km grid were not meaningful. The peak values were captured when applying resolutions of 500 m or smaller. Although slightly better results were obtained for 400 m, they chose 500 m grids as the appropriate spatial resolution considering the computational burden. Fully distributed models are more sensitive to resolution of the rainfall input as compared to semi-distributed models (Gires et al., 2015). Most of the current studies investigated either the effects of the model input resolution or the spin-up period on the model results. No study is known to the authors which explicitly assesses the effects of the spatial resolution of the model together with the length of the spin-up period and calibration period on the model performance.

We aim to comprehensively investigate the impact of the three major but overlooked pillars, (1) calibration period, (2) spin-up period and (3) spatial model resolution, on the calibration and validation performance of a physically-based distributed hydrological model for the Moselle River basin in France and Germany. The study area and data are introduced in section 2. The model and calibration framework are presented in section 3. The calibration and validation results are presented and discussed in sections 4 and 5. Finally, the key conclusions are drawn in Section 6.

2 Study Area and Data

2.1 Study area

The focus of this study is the Moselle River basin (Figure 1), i.e. the largest sub-basin of the Rhine River. The main channel of the Moselle River has a length of about 545 km (Demirel et al., 2013). The Moselle River basin, covering parts of the three countries France, Germany and Luxembourg, has a surface area of approximately 27262 km². The three longest tributaries of the Moselle River are the Saar, Sauer and Meurthe. The basin has different lithological and topographic characteristics, while it has a rain dominated regime (Brenot et al., 2007). The minimum, mean and maximum discharge values observed for the Moselle (at Cochem station) are 14 (dry summer), 130 (long term average until 2009) and 4000 m^3/s (winter), respectively (Demirel et al., 2013). The mean altitude of the basin is around 340 m and the land use is dominated by agriculture (54%) with arable areas, pastures and natural grasslands (Uehlinger et al., 2009), and forests (37%) in the mountains and hillslopes (Demirel et al., 2019).

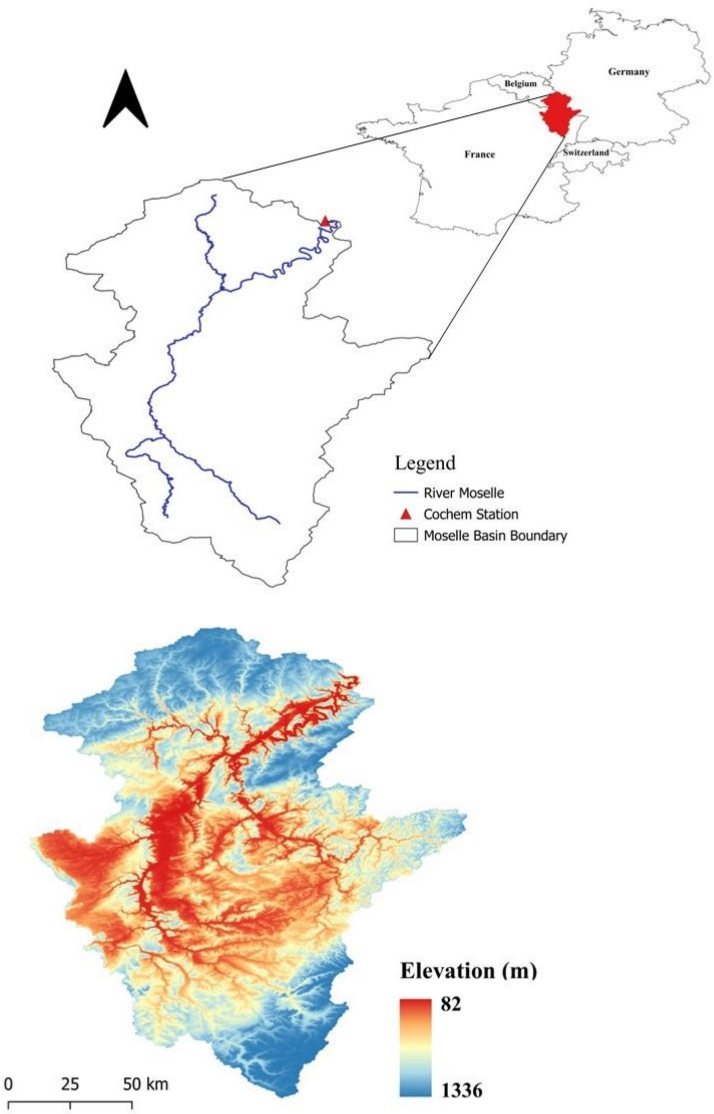

Figure 1. Moselle River network, basin boundary and elevation map

2.3 Data

Distributed hydrological models not only need hydrometeorological and geographical data as input, but also require parameters relevant for different hydrological processes such as interception and infiltration. At this point, the data availability and the spatio-temporal resolution of the input data play a vital role in the accuracy of a model. In this study, the model uses spatially distributed precipitation, temperature and potential evapotranspiration data as input

(Table 1). Meteorological data are from the E-OBS gridded data set on a daily basis (Haylock et al., 2008) and the discharge data at Cochem station was obtained from the Global Runoff Data Center (GRDC) in Koblenz (Germany).

The digital elevation model (DEM) is based on the Shuttle Radar Topography Mission (SRTM) from NASA (Ballabio et al., 2016). The soil classes are derived from the harmonized world soil database (Fischer et al., 2012), while land cover data is provided from the CORINE data set (Girard et al., 2019). Table 1 provides a brief summary of the data used in this study.

Table 1 Summary of geographical and meteorological data used as input for mHM.

Variable	**Description**	**Spatial Resolution**	**Temporal Resolution**	**Source**
Q (daily)	Streamflow (Cochem station, #6336050)	Point	Daily	GRDC
P (daily)	Precipitation	24 km	Daily	E-OBS 20.0e, MODIS
ET_{ref} (daily)	Reference evapotranspiration	24 km	Daily	E-OBS 20.0^e, MODIS
T_{avg} (daily)	Average air temperature	24 km	Daily	E-OBS 20.0^e, MODIS
Land cover	Pervious, impervious and forest	250 m	1 map for all period	CORINE
DEM data	Slope, aspect, flow accumulation and direction	250 m	1 map for all period	SRTM
Geology class	Two main geological formations	250 m	1 map for all period	EUROPEAN SOIL DATABASE
Soil class	Soil texture data	250 m	1 map for all period	HARMONIZED WORLD SOIL DATABASE

(SRTM: Shuttle Radar Topography Mission, CORINE: Coordination of Information on the Environment, GRDC: Global Runoff Data Center)

3 Methods

3.1 Meso-scale hydrological model

The grid-based meso-scale hydrological model (mHM) is a fully-distributed model in which for each grid cell incoming and outgoing fluxes for different storage compartments are calculated and the water balance of each compartment is updated after each time step (Dembélé et al., 2020; Kumar et al., 2013; Samaniego et al., 2010). In mHM, runoff is transferred to the downstream cells along the basin and river using three different routing methods i.e. Muskingum, adaptive time step with constant celerity and adaptive time step with varying celerity (Thober et al., 2019). In this study, we used adaptive time step with constant celerity method as it only requires one parameter i.e. streamflow celerity. In the last decade, mHM has been applied to basins in many countries in Europe (Marx et al., 2017; Samaniego et al., 2018), including Germany (Baroni et al., 2019; Höllering et al., 2018; Jing et al., 2019) and Denmark (Demirel et al., 2018; Koch et al., 2018), as well as to various large basins world-wide (Eisner et al., 2017; Huang et al., 2018).

mHM is an open source software written in the Fortran 2003 language and accessible from www.ufz.de/mhm, while the model is also compatible with many platforms, such as Linux, Mac and Windows (Nijssen et al., 2001; Samaniego et al., 2021). One of the most appealing features of the model code is the transferability between different input resolutions (Figure 2) for the desired computational resolutions (mesh). The model handles different resolutions of soil related data and meteorological data (Figure 2) by automatic upscaling and downscaling of high resolution geographical data (L0) and coarse meteorological data (L2) to reach the user-defined hydrological output resolution (L1). Also, the model provides flexibility to select a routing resolution (L11) different than the hydrological resolution (L1), so that the user can benefit from high resolution geographical input (soil, geology, aspect, LAI, elevation etc.) and does not loose time with preprocessing of meteorological data to fit the resolutions for model runs. Transferring data to a coarser resolution is done based on harmonic averaging instead of arithmetic averaging. In addition, different temporal resolutions for the model outputs can be used, e.g. daily, monthly or annual model results. For details of the process formulations, the readers and potential users may refer to the model papers (Kumar et al., 2013; Samaniego et al., 2010).

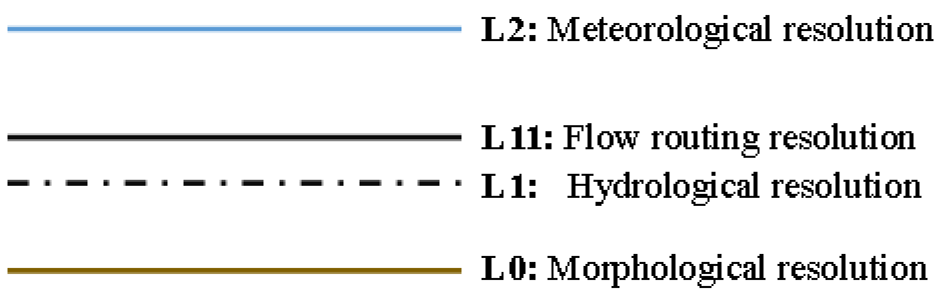

Figure 2. Model input and output scale configuration in mHM.

3.2 Parameter sensitivity analysis

Sensitivity analysis (SA) is an important step before calibration and validation of complex hydrological models to reduce the dimension of the search space. This will increase the effectiveness of the calibration process by reducing the run time. mHM includes around 55 global parameters used in physically based equations representing the different hydrological processes. In this study, we applied a local sensitivity analysis method based on the Jacobian matrix available in the PEST tool (Doherty, 2010). The parameters are perturbated one-at-a-time with a particular percentage (i.e. 5%) and the change in the performance metric is observed. PEST allows one side (only increase) or two side (increase and decrease) sensitivity analysis. We applied two side SA which required 2n+1 model runs (n is the number of parameters), i.e. 55 parameters x 2 sides + 1 =111 model runs.

3.3 Model calibration and validation

Since we are interested in capturing peak flows, we selected the Nash Sutcliffe Efficiency (NSE), i.e. the most commonly used metric in flood hydrology (Knoben et al., 2019), to present our calibration results. In this study, mHM version 5.10 was set-up for the Moselle River basin, and the effects of the three factors (pillars) on the model performance were examined. Accordingly, we tested all possible combinations of three factors, i.e. a total of 105 different cases comprising of three calibration data lengths, seven spin-up periods and five spatial model resolutions to design an appropriate calibration framework for the Moselle River basin. Here, we tested spatial model resolutions varying from 1 to 12 km (Figure 3). The mHM model internally upscales and downscales the input data to match the input scale to the hydrological model scale. Since we identified very small effects of the routing scale on the model performance, we fixed the routing scale to 6 km to save a substantial amount of run-time using the workstation configuration of the AMD Ryzen Threadripper 1900X 8-Core Processor (Win-10, 4.10 GHz and 64GB RAM). Further, we used three different calibration periods between 1991-2005, 1996-2005 and 2001-2005, corresponding to data lengths of 15, 10 and 5 years respectively. The four year period between 2006 and 2009 was selected as validation period for each model since we had data from 1991 to 2009. We tested seven spin-up period of 0, 1, 2, 3, 4, 5 and 10 years and five different spatial model resolutions of 1, 2, 4, 8 and 12 km. It should be noted that the geographical and geomorphological data of the mHM model is at a 250 m resolution and meteorological inputs (P, ET_{ref} and T_{avg}) are at a 24 km resolution. The discharge data at Cochem station was used both in the calibration and validation.

In addition, mHM internal auto-calibration tool provides four search algorithms. In this study, the Dynamically Dimensioned Search (DDS) algorithm (Tolson & Shoemaker, 2007) is used to calibrate the model parameters, since DDS is a fast converging method compared to local gradient based methods such as the steepest descent algorithm (Huot et al., 2019). Tolson and Shoemaker (2007) also highlighted that DDS outperformed one of the most popular optimization algorithm in hydrology i.e. Shuffled Complex Evolutionary algorithm (Duan et al., 1992). For a comprehensive analysis of the search space, we set the maximum number of iterations to 3000 model runs.

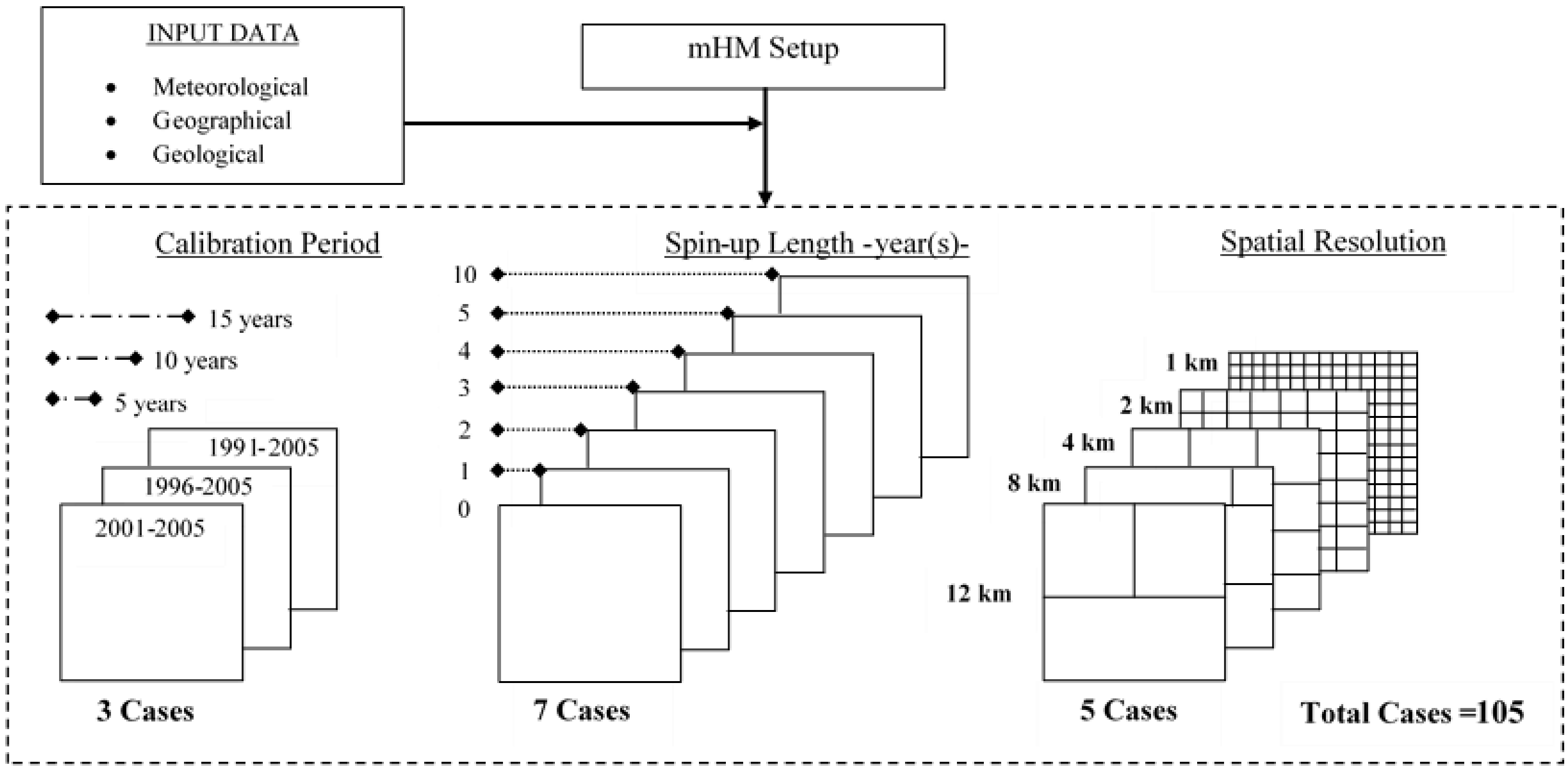

Figure 3. The framework of the study. Each of the 105 cases has been calibrated with the Dynamically Dimensioned Search (DDS) algorithm with a maximum number of 3000 iterations and the Nash-Sutcliffe Efficiency (NSE) as objective function.

4 Results

4.1 Parameter sensitivity analysis

Table 2 shows the most important 18 parameters using the NSE metric and sorted based on the normalized sensitivities. The normalized values are used to take into account both intial parameter values and raw sensitivity indicators from the Jacobian matrix. This is a more objective way as compared to using raw sensitivities directly, since a small change in some very small valued parameters may have a huge impact on the results whereas high valued geo-parameters may have a small raw sensitivity. In this approach, intial parameter values and raw sensitivities are multiplied (4th column) and then normalized by the maximum of this column. The normalized sensitivity value of the most sensitive parameter is 1 in this approach. Around two-third of the 55 parameters were not influencial on the streamflow dynamics and similar parameters found to be sensitive in other mHM studies in different basins (Demirel et al., 2018)

Table 2 Most sensitive parameters of mHM based on NSE performance.

Parameter	Initial value (-)	Raw sensitivity (-)	Abs (init. value* raw sensitivity) (-)	Normalized Sensitivity (-)
rotfrcoffore	0.9878	3.0199	2.9831	1.0000
rotfrcofclay	0.9637	1.8252	1.7590	0.5900
ptfksconst	-1.3251	0.4033	0.5344	0.1790
rotfrcofimp	0.9352	0.4676	0.4374	0.1470
ptflowconst	0.7518	0.3340	0.2511	0.0840
pet_bb	0.8942	0.2243	0.2006	0.0670
rechargcoef	6.4266	0.0260	0.1674	0.0560
pet_ap	0.4337	0.3569	0.1548	0.0520
ptfkssand	0.0094	16.2841	0.1527	0.0510
ptflowdb	-0.3323	0.4565	0.1517	0.0510
expslwintflw	0.0568	2.4514	0.1391	0.0470
pet_cc	-0.6204	0.1749	0.1085	0.0360
slwintreceks	13.3225	0.0077	0.1027	0.0340
pet_af	1.0445	0.0815	0.0851	0.0290

ptfksclay	0.0035	11.2824	0.0399	0.0130
thetanormc1	0.4722	0.0749	0.0354	0.0120
geoparam4	215.6520	0.0002	0.0335	0.0110
muskatrivslp	0.4657	0.0674	0.0314	0.0110

4.2 Effect of calibration data length on model performance

Figure 4 shows the model performance results in the calibration (left column) and validation (right column) periods as a function of the calibration data length for different spin-up periods and spatial resolutions. Besides the calibration results, we also present validation results as an independent test to evaluate the effects of the 105 cases.

For a 1 km resolution, Figure 4a shows that the model calibration performance varies depending on the spin-up period when the calibration data length increases from 5 to 15 years. Besides, Figure 4b indicates that the model validation performance increases with increasing calibration data length independently from the spin-up period. The results obtained for a 4 km resolution showed that the model calibration performance decreased when the calibration data length increased from 5 to 15 years except for a 1-year spin-up period (Figure 4c). However, the model validation performance increased when the calibration data length increased from 5 to 10 years and did not show a significant change between 10 and 15 years for 4 km resolution (Figure 4d). Figure 4e shows that the increase in calibration data length from 10 years to 15 years did not lead to significant changes in model calibration performance for a 8 km resolution except for a 0-year spin-up period. In addition, Figure 4f illustrates that the increase in calibration data length from 10 to 15 years deteriorates the model validation performance for spin-up periods of 4, 5 and 10 years. Overall, a calibration data length of 10 years is sufficient for 4-km and 8-km resolutions, whereas setting the calibration data length to 15 years is required when the spatial resolution of the model is 1 km.

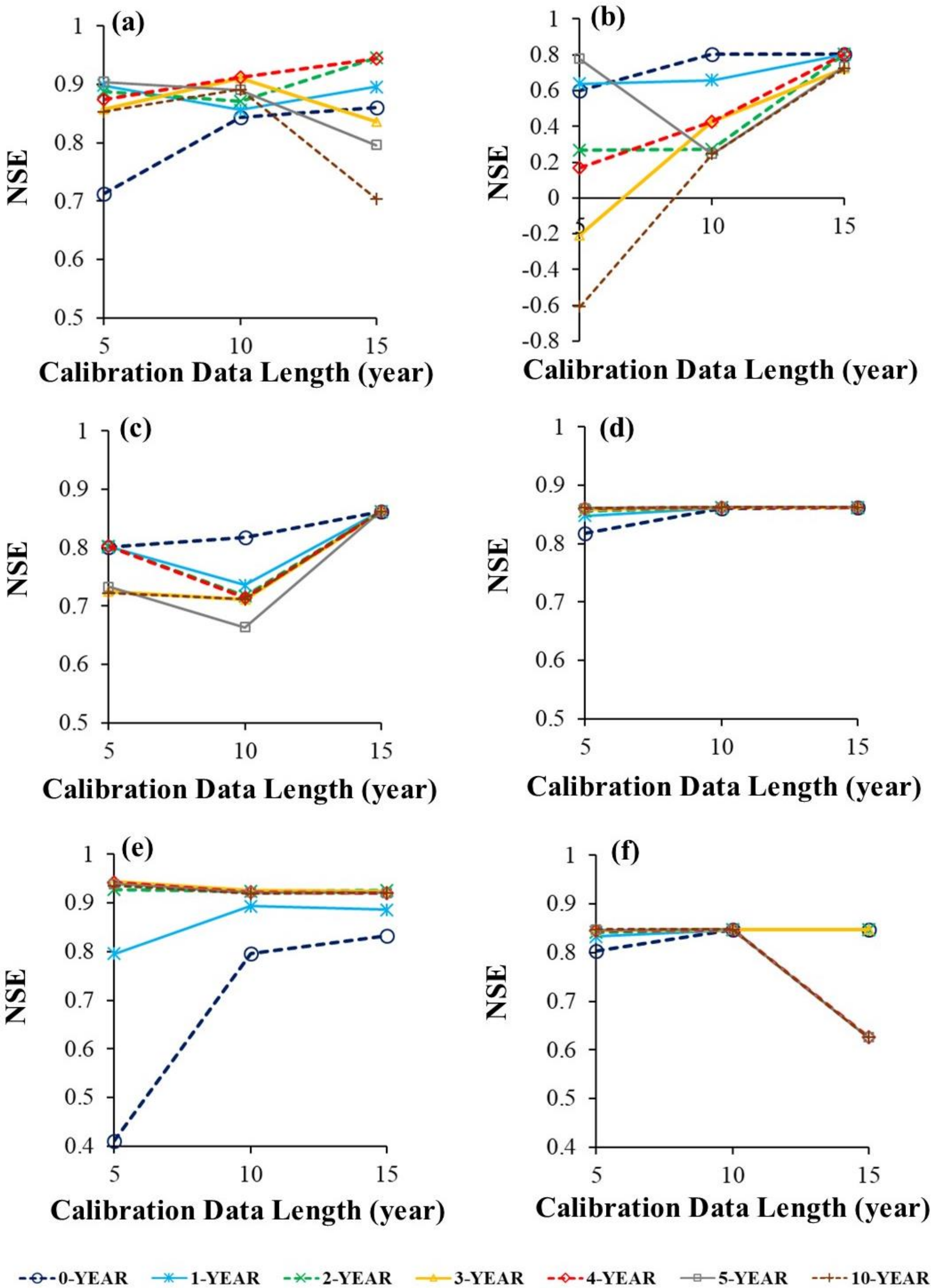

Figure 4. Model results (NSE) as a function of calibration period (length) for different spin-up periods (0 to 10 years) and different spatial resolutions; i.e. **a)** and **b)** 1 km **c)** and **d)** 4km, **e)** and

f) 8 km. Left column represents the calibration results; right column represents the validation results. Horizontal axis points to three scenarios i.e. 5 year calibration covers 2001-2005, 10 year calibration covers 1996-2005 and 15 year calibration covers 1991-2005.

4.3 Effect of spin-up period on model performance

Figure 5 highlights the impacts of the different spin-up periods on model performance by means of the NSE for different spatial resolutions and calibration data lengths. It is apparent from Figure 5 that an increase in spin-up period results in a higher model calibration performance (except the case with a calibration data length of 5 years and a 1 km resolution) as the model better adapts to the basin states. However, one can observe a decreasing trend in the validation performance when the spin-up period was set between 0-year and 5 years, particularly at a spatial resolution of 1 km and 2 km, while for a calibration data length of 15 years (Figure 5f), we see a similar behavior for almost each spatial resolution (except for a 4 km resolution). Interestingly, for a calibration length of 15 years, from meso to coarse spatial model resolution (from 4 to 12 km), the model calibration performance jumps from a NSE value of 0.4 to 0.9 as the spin-up period increases from zero to two years (Figure 5e). With a few exceptions, model calibration and validation results show less sensitivity to changing spin-up periods after two years. On the other hand, the model calibration performance with a 1 and 2 km resolution show high sensitivity to the spin-up period. This is a clear indication of the importance of selecting an appropriate spin-up period for a selected spatial resolution in a systematic model calibration framework. In summary, considering a calibration data length of 10 years, a spin-up period of 2 years is found to be adequate for the application of mHM to the Moselle River basin.

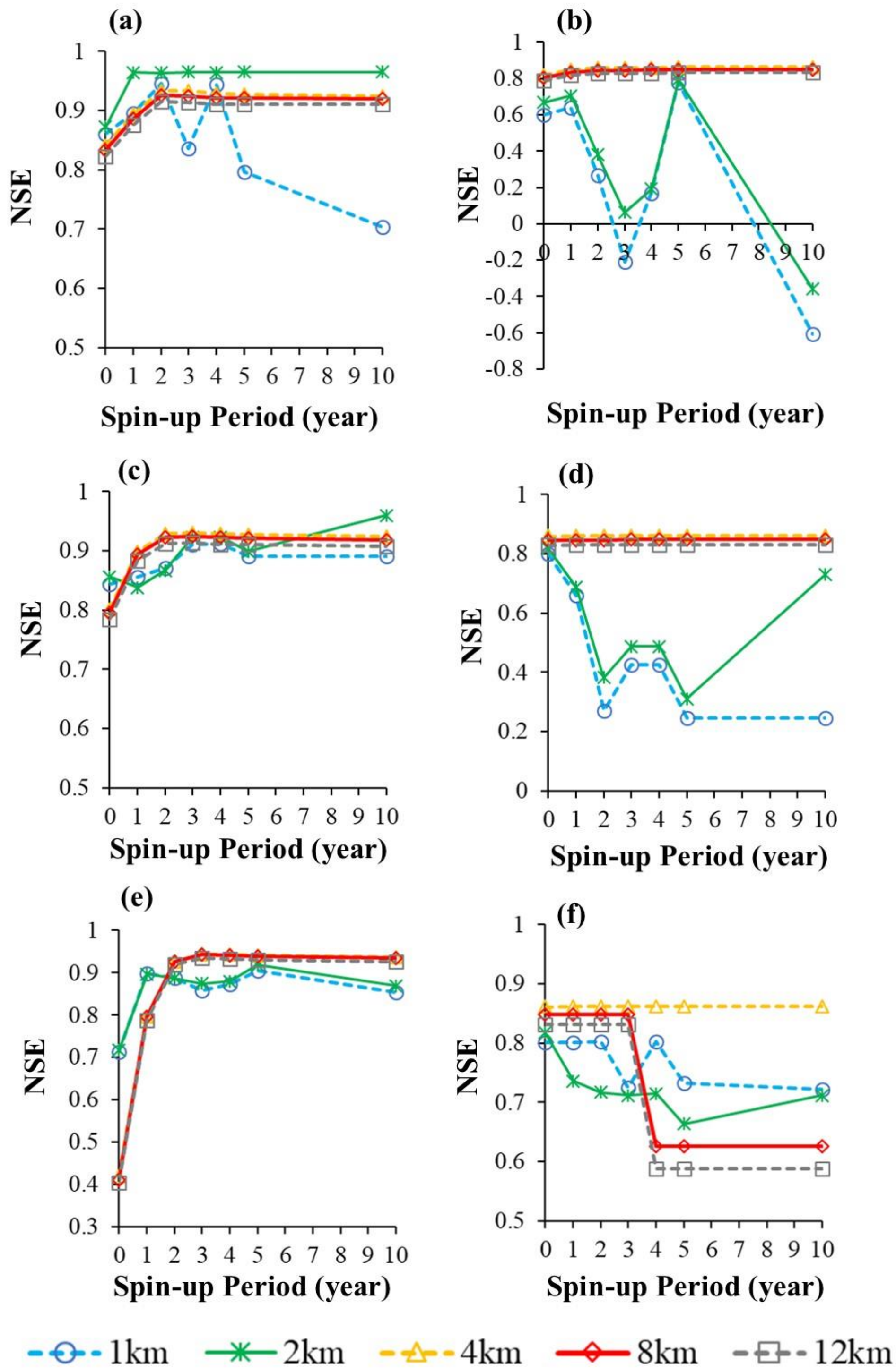

Figure 5. Model results (NSE) as a function of spin-up period for different spatial resolutions (1 km, 2km, 4km, 8km and 12 km) and calibration periods, i.e. **a)** and **b)** 5 years **c)**

and **d)** 10 years, **e)** and **f)** 15 years. Left column represents the calibration results; right column represents the validation results.

4.4 Effect of spatial resolution on model performance

Figure 6 shows the variation in NSE in the model calibration and validation as a function of spatial resolution. Colored lines represent different spin-up periods. Two adjacent sub-plots in each raw illustrates 5, 10 and 15 years of calibration data lengths, respectively. The results obtained for both model calibration and validation illustrated that the model performance increased as the model resolution increases from 1 to 4 km (except for the validation performance of 15 years calibration data length). Even though this is contrary to the expectations considering the physical point of view, this can be from the fact that different uncertainties in the input data are less influential (reduced) after averaging data to coarser scales (upscaling). In addition, Figure 6a depicts that a 2 km spatial resolution gave satisfactory results in model calibration, while the model shows the best validation performance when the spatial resolution is set to 4 km (Figure 6b). Also, for a calibration data length of 10 years, a 4 km resolution seems the best option for both calibration and validation (Figure 6c and Figure 6d). However, some inconsistencies may exist for shorter spin-up periods (such as a 0-year spin-up period). What is striking about the cases with a 15-year calibration period is that there is no improvement in model performance beyond a spatial resolution of 4 km (Figure 6e and Figure 6f) as the NSE values tend to decrease towards 8 and 12 km resolutions.

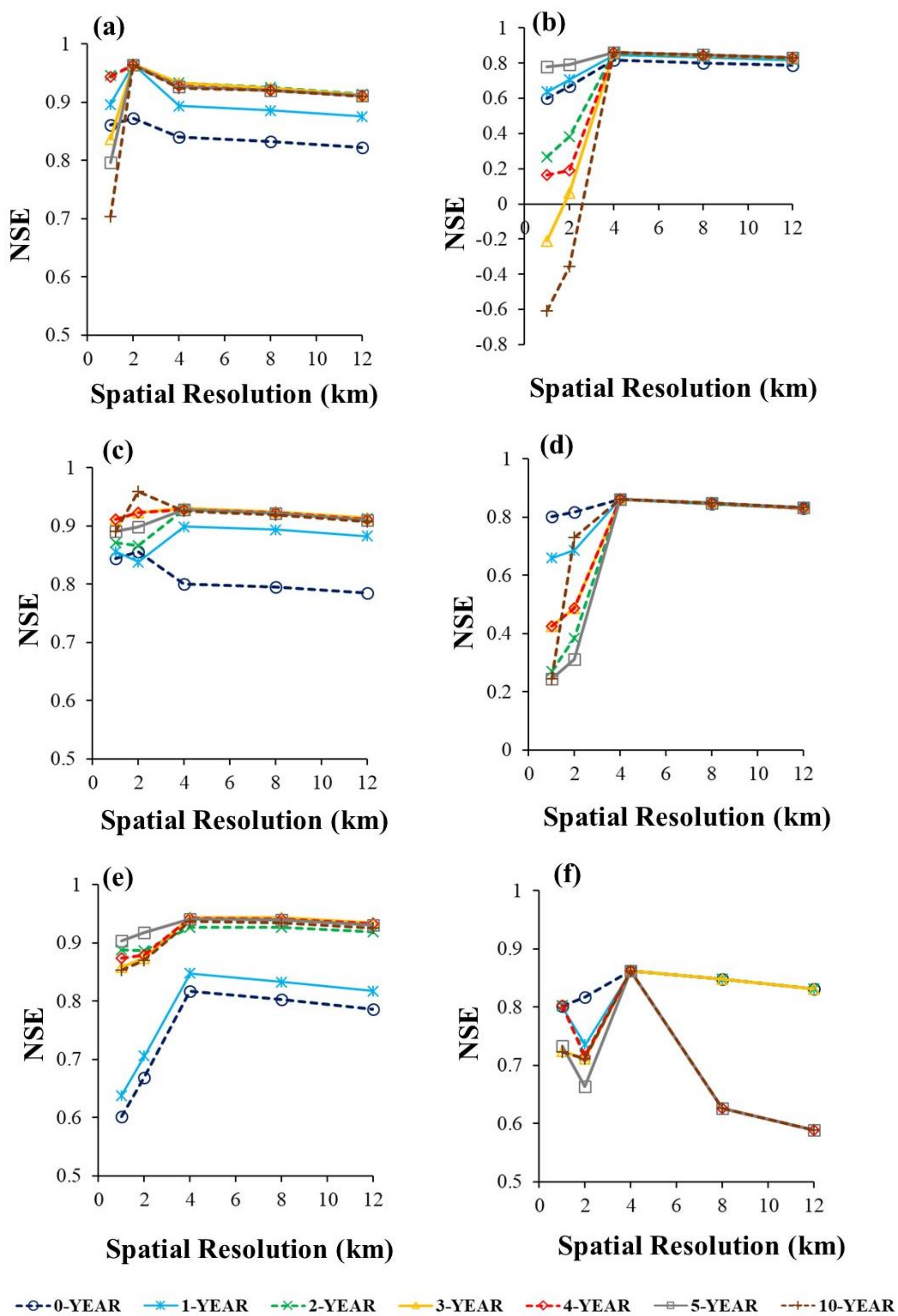

Figure 6. Model results (NSE) as a function of spatial resolution for different spin-up periods (0 to 10 years) and calibration periods, i.e. **a)** and **b)** 5 years from 2001 to 2005 **c)** and **d)** 10 years from 1996 to 2005, **e)** and **f)** 15 years from 1991 to 2005. Left column represents the calibration results; right column represents the validation results.

5 Discussion

Model calibration

Model calibration is usually executed with the available data and computational resources. More data and higher model resolutions are assumed to provide a more realistic simulation requiring less need for model calibration than those with coarser data. In this study, we analyzed 105 different model calibrations to identify an appropriate configuration of three pillars, i.e. calibration data length, spin-up period and spatial resolution. We followed a smart sampling approach for the choice of experimental details. For instance, instead of testing all spin-up periods from one year to ten years, we only focused on zero to five years with one year interval and added an experiment with a ten year spin-up period as the last case. Similarly, we included only some of the most commonly used spatial model resolutions, i.e. 1, 2 and 4 km. Although we could include more spatial resolutions between 250 m (L0 geographical data resolution) and 24 km (L2 meteorological data resolution) such as 3, 6 and 24 km, we only considered two additional resolutions (8 and 12 km). Testing 11 spin-up periods (i.e. 0 to 10 years) together with 10 spatial resolutions (i.e. 250 m, 500 m, 1, 2, 3, 4, 6, 8, 12 and 24 km) would enormously increase the number of cases directly affecting the total duration of the calibration experiments. This would also raise the question of redundancy due to the testing of minor changes in the resolutions and spin-up periods. Furthermore, the model is incapable of upscaling and downscaling of model inputs for the non-integer spatial resolutions, e.g. 5, 7, 9, 10, 11 and 23 km.

Although NSE is the most commonly used metric to assess hydrological model performance (Mizukami et al., 2019), it is critisized for being dominated by high flow performance (Pushpalatha et al., 2012). We used the DDS method, which is available in the model tool, to calibrate our model. To develop a full picture of hydrological model behavior, additional studies will be needed that consider multi-objective calibrations using pareto archived DDS (Asadzadeh & Tolson, 2009) with additional metrics such as the Kling-Gupta Efficiency (Gupta et al., 2009) and Spatial Efficiency (Demirel et al., 2018). We chose a sufficiently large number of iterations (3000 runs) and reached reasonable performance results. Here, our motivation was to scan a wide spectrum of the parameter domain instead of a short calibration with several hundereds of iterations. Also, we only focused on single gage temporal calibration with NSE. Further research should investigate effect of multi-gage and spatial model calibrations using Spatial Efficiency (SPAEF) as objective function to assess the model performance (Demirel et al., 2018).

Effect of three pillars on model performance

Based on the trade-off between available data and computational resources, the modeler has to choose an appropriate combination of the three pillars. In this study, we assessed the effect of each pillar on the model performance. It is somewhat surprising that higher spatial model resolutions (1 and 2 km) lead to a higher sensitivity to the length of the calibration period. For spin-up periods longer than 2 years, the model performance is relatively less sensitive. This indicates that using a longer spin-up period in hydrological simulations does not always have a positive effect on the model performance. From a physical point of view, the spin-up period should be basin dependent and influenced by factors such as geographical heterogenity, land cover anduse and flow regime. For instance, in rainfed catchments, the performance of hydrological models is relatively higher than those in snowmelt dominated regions which can reduce the dependency of the model for longer data length and spin-up period. However, capturing rainfall heterogeneity at higher spatial resolutions is necessary for better performance. The size of the catchment (Wallace et al., 2018), heterogeneity of rainfall (Nicótina et al., 2008) and karstic geomorphology can greatly effect the spatio-temporal variations of hydrological processes and three pillars (Zhang et al., 2020). Larger gird-size (coarser spatial resolution) can be used in larger basins whereas especially for the latter cases (rainfall heterogeneity and complex geology), the need for better quality data and longer time series increases significantly. We are aware that spin-up periods longer than 5 years are not realistic in many hydrological modeling studies (Ajami et al. 2014), however, we intended to test a wide range of periods.

Spatial model resolution directly effects the number of cells and the pattern of the hydrological variable, e.g. actual evapotranspiration (AET), over the model domain (Booij, 2002; Chen et al., 2017; Cosgrove et al., 2003; Etchevers et al., 2001; Zheng et al., 2018). For instance, a single cell with spatial resolution of 24 km does not provide any pattern of AET depending on the vegetation and soil type. To have a descent histogram of the spatial patterns, resolutions that result in around 1000-2000 cells (pixels) are required to calculate spatial performance as shown in other basins (Demirel et al., 2018).

Uncertainties and Data

Assessing uncertainities raising from model structure, inputs and parameters is important for assessing the reliability of the results. Model structure uncertainty can be analyzed by using multiple models (Demirel et al., 2013). Here, we only focused on one distributed model (i.e. mHM) and the EOBS meteorological dataset. Parameter uncertainty is assumed to be reduced during the model calibration. There are still many unanswered questions about the model input uncertainty. To compare the effect of input uncertainty on the results, the ERA5 meteorological dataset (Hersbach et al., 2020) can be used in the model in addition to the EOBS dataset (Cornes et al., 2018). Further, we chose aspect based potential ET correction in the model as leaf area index (LAI) based potential ET correction will be a topic of our future study. It is assumed that the LAI based potential ET correction would yield better AET estimates; therefore, better discharge performance as compared to those with aspect data (Demirel et al., 2018).

Data quality and length can be big issues for modelers from developing countries. Even though the modeler has a long time series with unlimited computational resources, a ten-year part of the new data set with a spin-up period of two or three years is sufficient for the model calibration. Then, the remaining, i.e. not wasted, data can be used for model validation (Royer-Gaspard et al., 2021). Further work should examine the effect of model input data resolution in addition to the model spatial resolution. Also, the length of the validation period can be varied in addition to the length of the calibration period.

6 Conclusions

This study was designed to comprehensively investigate the effects of three user-defined model configurations that are usually determined based on local expert knowledge and available data. We focused on the identification of the appropriate length of the calibration period, the length of the spin-up period and the appropriate spatial model resolution for the Moselle River basin. For that, we used a fully distributed hydrological model (mHM) and performed 105 different calibrations with the DDS optimization algorithm and NSE objective function. The 105 cases are combinations of three calibration periods, seven spin-up periods and five spatial model resolutions.

The main conclusions from this work can be summarized as follows:

- Based on the results of the comparison of three calibration data lengths, 10 years is found to be an appropriate length for the Moselle River basin. The interaction between calibration period and 1-2 km spatial resolution has the strongest effect on the results.
- Based on the results of the comparison of three spin-up periods, two years of spin-up period in addition to the 10 years of calibration data is found to be sufficient for the model to adopt to the initial conditions in the Moselle River basin. Longer spin-up periods than two years did not significantly improve the model calibration and validation performances.
- Based on the results of the comparison of five spatial resolutions, 4 km is found to be the most appropriate model resolution for the Moselle River basin since the performance slightly deteriorated at coarser resolutions (i.e. 8 and 12 km).

Overall, the three factors analyzed in our study are usually overlooked in hydrological modeling. However, the results showed that we should carefully analyze the different combinations of calibration data length, spin-up period and spatial resolution instead of selecting an arbitrary combination. It is important to mention that our multi-case analysis framework proposed in this study can be applied to any other spatially distributed model and catchment.

Acknowledgments, Samples, and Data

We acknowledge the financial support for the SPACE project by the Villum Foundation (http://villumfonden.dk/) through their Young Investigator Programme (grant VKR023443). The second author (MCD) is supported by the National Center for High Performance Computing of Turkey (UHeM) under grant number 1007292019 and Ir. Cornelis Lely Stichting (grant 20957310). The mHM v5.10 model (available at https://git.ufz.de/mhm) together with all the set-up data was provided by UFZ in Leipzig, Germany.

References

Abdo, K. S., Fiseha, B. M., Rientjes, T. H. M., Gieske, A. S. M., & Haile, A. T. (2009). Assessment of climate change impacts on the hydrology of Gilgel Abay catchment in Lake Tana Basin, Ethiopia. *Hydrological Processes*, *2274*(November 2008), n/a-n/a. https://doi.org/10.1002/hyp.7363

Ajami, H., Evans, J. P., McCabe, M. F., & Stisen, S. (2014). Technical Note: Reducing the spin-up time of integrated surface water–groundwater models. *Hydrology and Earth System Sciences*, *18*(12), 5169–5179. https://doi.org/10.5194/hess-18-5169-2014

Ajami, Hoori, McCabe, M. F., Evans, J. P., & Stisen, S. (2014). Assessing the impact of model spin-up on surface water-groundwater interactions using an integrated hydrologic model. *Water Resources Research*, *50*(3), 2636–2656. https://doi.org/10.1002/2013WR014258

Andersen, J., Refsgaard, J. C., & Jensen, K. H. (2001). Distributed hydrological modelling of the Senegal River Basin - Model construction and validation. *Journal of Hydrology*, *247*(3–4), 200–214. https://doi.org/10.1016/S0022-1694(01)00384-5

Asadzadeh, M., & Tolson, B. A. (2009). A new multi-objective algorithm, pareto archived DDS. In *Proceedings of the 11th annual conference companion on Genetic and evolutionary computation conference - GECCO '09* (p. 1963). New York, New York, USA: ACM Press. https://doi.org/10.1145/1570256.1570259

Ashraf, A. (2013). Changing Hydrology of the Himalayan Watershed. In *Current Perspectives in Contaminant Hydrology and Water Resources Sustainability* (Vol. i, p. 13). InTech. https://doi.org/10.5772/54492

Ballabio, C., Panagos, P., & Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, *261*, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006

Baroni, G., Schalge, B., Rakovec, O., Kumar, R., Schüler, L., Samaniego, L., et al. (2019). A Comprehensive Distributed Hydrological Modeling Intercomparison to Support Process Representation and Data Collection Strategies. *Water Resources Research*, 990–1010. https://doi.org/10.1029/2018WR023941

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., et al. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290. Retrieved from http://dx.doi.org/10.1002/hyp.3360090305

Booij, M. J. (2002). Modelling the effects of spatial and temporal resolution of rainfall and basin model on extreme river discharge. *Hydrological Sciences Journal*, *47*(2), 307–320. https://doi.org/10.1080/02626660209492932

Booij, M. J. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1–4), 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Brenot, A., Carignan, J., France-Lanord, C., & Benoît, M. (2007). Geological and land use control on δ34S and δ18O of river dissolved sulfate: The Moselle river basin, France. *Chemical Geology*, *244*(1–2), 25–41. https://doi.org/10.1016/j.chemgeo.2007.06.003

Bucchignani, E., Mercogliano, P., Rianna, G., & Panitz, H.-J. (2016). Analysis of ERA-Interim-driven COSMO-CLM simulations over Middle East - North Africa domain at different spatial resolutions. *International Journal of Climatology*, *36*(9), 3346–3369. https://doi.org/10.1002/joc.4559

Chen, Y., Li, J., Wang, H., Qin, J., & Dong, L. (2017). Large-watershed flood forecasting with high-resolution distributed hydrological model. *Hydrology and Earth System Sciences*, *21*(2), 735–749. https://doi.org/10.5194/hess-21-735-2017

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017JD028200

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., et al. (2003). Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS). *Journal of Geophysical Research D: Atmospheres*, *108*(22). https://doi.org/10.1029/2002jd003316

Cuo, L., Giambelluca, T. W., Ziegler, A. D., & Nullet, M. A. (2006). Use of the distributed hydrology soil vegetation model to study road effects on hydrological processes in Pang Khum Experimental Watershed, northern Thailand. *Forest Ecology and Management*, *224*(1–2), 81–94. https://doi.org/10.1016/j.foreco.2005.12.009

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K. R., Zeckoski, R. W., Jeong, J., et al. (2015). A recommended calibration and validation strategy for hydrologic and water quality models. *Transactions of the ASABE*, *58*(6), 1705–1719. https://doi.org/10.13031/trans.58.10712

Dehotin, J., & Braud, I. (2008). Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. *Hydrology and Earth System Sciences*, *12*(3), 769–796. https://doi.org/10.5194/hess-12-769-2008

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR026085

Demirel, Özen, Orta, Toker, Demir, Ekmekcioğlu, et al. (2019). Additional Value of Using Satellite-Based Soil Moisture and Two Sources of Groundwater Data for Hydrological Model Calibration. *Water*, *11*(10), 2083. https://doi.org/10.3390/w11102083

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resources Research*, *49*(7), 4035–4053. https://doi.org/10.1002/wrcr.20294

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrol. Earth Syst. Sci.*, *22*(2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018

Doherty, J. (2010). *PEST: Model-Independent Parameter Estimation, User Manual: 5th Edition*. *PEST Manual*.

Duan, Q. Q.-Y. Y., Sorooshian, S., & Gupta, V. Effective and efficient global optimization for conceptual rainfall-runoff models, 28 Water Resources Research § (1992). https://doi.org/10.1029/91WR02985

Eisner, S., Flörke, M., Chamorro, A., Daggupati, P., Donnelly, C., Huang, J., et al. (2017). An ensemble analysis of climate change impacts on streamflow seasonality across 11 large river basins. *Climatic Change*, *141*(3), 401–417. https://doi.org/10.1007/s10584-016-1844-5

Epstein, D. J., Welles, E., & Day, G. N. (1998). Probabilistic Hydrologic Forecasting Methods and Tools. In *Proceedings of the First Federal Interagency Hydrologic Modeling Conference* (pp. 6(17-24)). Las Vegas, Nevada: Subcommittee on Hydrology of the Interagency Advisory Committee on Water Data.

Etchevers, P., Durand, Y., Habets, F., Martin, E., & Noilhan, J. (2001). Impact of spatial resolution on the hydrological simulation of the Durance high-Alpine catchment, France. *Annals of Glaciology*, *32*, 87–92. https://doi.org/10.3189/172756401781819337

Fischer, G., Nachtergaele, F., Prieler, S., Velthuizen, H. T. van, Verelst, L., & Wiberg, D. (2012). *Global Agro-ecological Zones–Model Documentation (GAEZ v. 3.0)*. *Food and Agriculture Organization of the United Nations*.

Gelleszun, M., Kreye, P., & Meon, G. (2017). Representative parameter estimation for hydrological models using a lexicographic calibration strategy. *Journal of Hydrology*, *553*, 722–734. https://doi.org/10.1016/j.jhydrol.2017.08.015

Girard, M.-C., Girard, C., Courault, D., Gilliot, J.-M., Loubersac, L., Meyer-Roux, J., et al. (2019). Corine Land Cover. In *Processing of Remote Sensing Data*. https://doi.org/10.1201/9780203741917-19

Gires, A., Giangola-Murzyn, A., Abbes, J. B., Tchiguirinskaia, I., Schertzer, D., & Lovejoy, S. (2015). Impacts of small scale rainfall variability in urban areas: a case study with 1D and 1D/2D hydrological models in a multifractal framework. *Urban Water Journal*, *12*(8), 607–617. https://doi.org/10.1080/1573062X.2014.923917

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *Journal of Geophysical Research Atmospheres*, *113*(20). https://doi.org/10.1029/2008JD010201

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., et al. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology*, *556*(1), 39–49. https://doi.org/10.1016/j.jhydrol.2017.11.009

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Höllering, S., Wienhöfer, J., Ihringer, J., Samaniego, L., & Zehe, E. (2018). Regional analysis of parameter sensitivity for simulation of streamflow and hydrological fingerprints. *Hydrology and Earth System Sciences*, *22*(1), 203–220. https://doi.org/10.5194/hess-22-203-2018

Huang, S., Kumar, R., Rakovec, O., Aich, V., Wang, X., Samaniego, L., et al. (2018). Multimodel assessment of flood characteristics in four large river basins at global warming of 1.5, 2.0 and 3.0 K above the pre-industrial level. *Environmental Research Letters*, *13*(12). https://doi.org/10.1088/1748-9326/aae94b

Huot, P. L., Poulin, A., Audet, C., & Alarie, S. (2019). A hybrid optimization approach for efficient calibration of computationally intensive hydrological models. *Hydrological Sciences Journal*, *64*(10), 1204–1222. https://doi.org/10.1080/02626667.2019.1624922

Jing, M., Heße, F., Kumar, R., Kolditz, O., Kalbacher, T., & Attinger, S. (2019). Influence of input and parameter uncertainty on the prediction of catchment-scale groundwater travel time distributions. *Hydrology and Earth System Sciences*, *23*(1), 171–190. https://doi.org/10.5194/hess-23-171-2019

Kim, K. B., Kwon, H. H., & Han, D. (2018). Exploration of warm-up period in conceptual hydrological modelling. *Journal of Hydrology*, *556*, 194–210. https://doi.org/10.1016/j.jhydrol.2017.11.015

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873–1886. https://doi.org/10.5194/gmd-11-1873-2018

Koren, V. I., Finnerty, B. D., Schaake, J. C., Smith, M. B., Seo, D. J., & Duan, Q. Y. (1999). Scale dependencies of hydrologic models to spatial variability of precipitation. *Journal of Hydrology*, *217*(3–4), 285–302. https://doi.org/10.1016/S0022-1694(98)00231-5

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*(1), 360–379. https://doi.org/10.1029/2012WR012195

Lohmann, D., Raschke, E., Nijssen, B., & Lettenmaier, D. P. (1998). Hydrologie à l'échelle régionale: II. Application du modèle VIC-2L sur la rivière Weser, Allemagne. *Hydrological Sciences Journal*, *43*(1), 143–158. https://doi.org/10.1080/02626669809492108

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, *26*(2), 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1

Marx, A., Kumar, R., Thober, S., Zink, M., Wanders, N., Wood, E. F., et al. (2017). Climate change alters low flows in Europe under a 1.5, 2, and 3 degree global warming. *Hydrology and Earth System Sciences Discussions*, 1–24. https://doi.org/10.5194/hess-2017-485

Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water Resources Research*, *45*(9), W09405. https://doi.org/10.1029/2009wr007872

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrology and Earth System Sciences*, *23*(6), 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

Näschen, K., Diekkrüger, B., Leemhuis, C., Steinbach, S., Seregina, L., Thonfeld, F., & van der Linden, R. (2018). Hydrological Modeling in Data-Scarce Catchments: The Kilombero Floodplain in Tanzania. *Water*, *10*(5), 599. https://doi.org/10.3390/w10050599

Nicótina, L., Alessi Celegon, E., Rinaldo, A., & Marani, M. (2008). On the impact of rainfall patterns on the hydrologic response. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006654

Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F. (2001). Predicting the discharge of global rivers. *Journal of Climate*, *14*(15), 3307–3323.

Pang, J., Zhang, H., Xu, Q., Wang, Y., Wang, Y., Zhang, O., & Hao, J. (2020). Hydrological evaluation of open-access precipitation data using SWAT at multiple temporal and spatial scales. *Hydrology and Earth System Sciences*, *24*(7), 3603–3626. https://doi.org/10.5194/hess-24-3603-2020

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., & Mathevet, T. (2007). Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, *52*(1), 131–151. https://doi.org/10.1623/hysj.52.1.131

Pushpalatha, R., Perrin, C., Moine, N. Le, & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420–421*, 171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055

Rahman, M. M., Lu, M., & Kyi, K. H. (2016). Seasonality of hydrological model spin-up time: a case study using the Xinanjiang model. *Hydrology and Earth System Sciences Discussions*, 1–22. https://doi.org/10.5194/hess-2016-316

Raihan, F., Beaumont, L. J., Maina, J., Saiful Islam, A., & Harrison, S. P. (2020). Simulating streamflow in the Upper Halda Basin of southeastern Bangladesh using SWAT model. *Hydrological Sciences Journal*, *65*(1), 138–151. https://doi.org/10.1080/02626667.2019.1682149

Revilla-Romero, B., Wanders, N., Burek, P., Salamon, P., & de Roo, A. (2016). Integrating remotely sensed surface water extent into continental scale hydrology. *Journal of Hydrology*, *543*, 659–670. https://doi.org/10.1016/j.jhydrol.2016.10.041

Rodell, M., Houser, P. R., Berg, A. A., & Famiglietti, J. S. (2005). Evaluation of 10 methods for initializing a land surface model. *Journal of Hydrometeorology*, *6*(2), 146–155. https://doi.org/10.1175/JHM414.1

Royer-Gaspard, P., Andréassian, V., & Guillaume, T. (2021). Technical note: PMR – a proxy metric to assess hydrological model robustness in a changing climate. *Hydrol. Earth Syst. Sci. Discuss.*, *preprint*. https://doi.org/10.5194/hess-2021-58

Samaniego, Brenner, J., Craven, J., Cuntz, M., Dalmasso, G., Demirel, M. C., et al. (2021, February 3). mesoscale Hydrologic Model - mHM v5.11.0. Leipzig. https://doi.org/10.5281/ZENODO.4462822

Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., et al. (2018). Anthropogenic warming exacerbates European soil moisture droughts. *Nature Climate Change*, *8*(5), 421–426. https://doi.org/10.1038/s41558-018-0138-5

Samaniego, Luis, Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5), W05523. https://doi.org/10.1029/2008WR007327

Shi, X., Wood, A. W., & Lettenmaier, D. P. (2008). How essentialis hydrologic model calibration to seasonal stream flow forecasting? *Journal of Hydrometeorology*, *9*(6), 1350–1363. https://doi.org/10.1175/2008JHM1001.1

Shrestha, R., & Houser, P. (2010). A heterogeneous land surface model initialization study. *Journal of Geophysical Research Atmospheres*, *115*(19), 1–10. https://doi.org/10.1029/2009JD013252

Simunek, J. J., Genuchten, M. T. Van, & Šejna, M. (2012). HYDRUS: Model Use, Calibration, and Validation. *Transactions of the ASABE*, *55*(4), 1263–1276. https://doi.org/10.13031/2013.42239

Sood, A., & Smakhtin, V. (2015). Revue des modèles hydrologiques globaux. *Hydrological Sciences Journal*, *60*(4), 549–565. https://doi.org/10.1080/02626667.2014.950580

Sood, A., Muthuwatta, L., & McCartney, M. (2013). A SWAT evaluation of the effect of climate change on the hydrology of the Volta River basin. *Water International*, *38*(3), 297–311. https://doi.org/10.1080/02508060.2013.792404

Sorooshian, S., Gupta, H. V., & Rodda, J. C. (1997). *Land Surface Processes in Hydrology:Trials and Tribulations of Modeling and Measuring*. (S. Sorooshian, H. V. Gupta, & J. C. Rodda, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-60567-3

Sreedevi, S., & Eldho, T. I. (2019). A two-stage sensitivity analysis for parameter identification and calibration of a physically-based distributed model in a river basin. *Hydrological Sciences Journal*, *64*(6), 701–719. https://doi.org/10.1080/02626667.2019.1602730

Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., & Samaniego, L. (2019). The multiscale routing model mRM v1.0: simple river routing at resolutions from 1 to 50 km. *Geoscientific Model Development*, *12*(6), 2501–2521. https://doi.org/10.5194/gmd-12-2501-2019

Tian, J., Liu, J., Wang, Y., Wang, W., Li, C., & Hu, C. (2020). A coupled atmospheric-hydrologic modeling system with variable grid sizes for rainfall-runoff simulation in semi-humid and semi-arid watersheds: How does the coupling scale affects the results? *Hydrology and Earth System Sciences Discussions*, 1–36. https://doi.org/10.5194/hess-2019-587

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Uehlinger, U., Wantzen, K. M., Leuven, R. S. E. W., & Arndt, H. (2009). The Rhine River Basin. In *Rivers of Europe* (pp. 199–245). London: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/B8M9H-4VJ0XHB-R/2/5a2ee83572d773dd7ec4e308c6a3c292

Wallace, C., Flanagan, D., & Engel, B. (2018). Evaluating the Effects of Watershed Size on SWAT Calibration. *Water*, *10*(7), 898. https://doi.org/10.3390/w10070898

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., & Seibert, J. (2020). Hydrological model calibration with uncertain discharge data. *Hydrological Sciences Journal*, *00*(00), 1–16. https://doi.org/10.1080/02626667.2020.1735638

Xu, Y. P., Zhang, X., Ran, Q., & Tian, Y. (2013). Impact of climate change on hydrology of upper reaches of Qiantang River Basin, East China. *Journal of Hydrology*, *483*, 51–60. https://doi.org/10.1016/j.jhydrol.2013.01.004

Yang, Z. L., Dickinson, R. E., Henderson-Selles, A., & Pitman, A. J. (1995). Preliminary study of spin-up processes in land surface models with the first stage data of project for intercomparison of land surface parameterization schemes phase 1(a). *Journal of Geophysical Research*, *100*(D8). https://doi.org/10.1029/95jd01076

Zhang, Z., Chen, X., Cheng, Q., Li, S., Yue, F., Peng, T., et al. (2020). Coupled hydrological and biogeochemical modelling of nitrogen transport in the karst critical zone. *Science of The Total Environment*, *732*, 138902.

https://doi.org/10.1016/j.scitotenv.2020.138902

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., & Zhang, T. (2018). On Lack of Robustness in Hydrological Model Development Due to Absence of Guidelines for Selecting Calibration and Evaluation Data: Demonstration for Data-Driven Models. *Water Resources Research*, *54*(2), 1013–1030. https://doi.org/10.1002/2017WR021470