~~Regionalization~~ Regionalization in global hydrological models and its impact on ~~global~~ runoff simulations: A case study using ~~the global hydrological model~~ WaterGAP3 (v 1.0.0)

Jenny Kupzig[1], Nina Kupzig[2], Martina Flörke[1]

[1]Institute of Engineering Hydrology and Water Resources Management, Ruhr-University, 44801, Bochum, Germany
[2]Faculty of Management and Economics, Ruhr-University, 44780, Bochum, Germany

Correspondence to: Jenny Kupzig (jenny.kupzig@rub.de)

Abstract:

Valid simulation results from global hydrological models (GHMs), such as WaterGAP3, are essential to detecting hotspots or studying patterns in climate change impacts. However, the lack of worldwide monitoring data makes it challenging to adapt GHMs' parameters to enable such valid simulations globally. Therefore, ~~regionalization~~ regionalization is necessary to estimate parameters in ungauged basins. This study presents the results of ~~new regionalization~~ regionalization methods for the first time applied on the GHM WaterGAP3~~ and~~. It aims to provide insights into (1) selecting a suitable ~~regionalization~~ regionalization method ~~method~~ and (2) evaluating its impact on ~~the~~ runoff simulation. In this study, ~~Our results suggest that machine learning-based methods may be too flexible for regionalizing WaterGAP3 due to a significant performance loss between training and testing.~~ four regionalization methods have been identified as appropriate for WaterGAP3. These methods span the full spectrum of methodologies, i.e., regression-based methods, physical similarity, and spatial proximity, using traditional and machine learning-based approaches. Moreover, the methods differ in the descriptors used to achieve optimal results, although all utilize climatic and physiographic descriptors. This demonstrates (1) that different methods use descriptor sets with varying efficiency and (2) that combining climatic and physiographic descriptors is optimal for regionalizing worldwide basins. ~~In contrast, the most basic regionalization method (using the concept of spatial proximity) outperforms most of the developed regionalization methods and a pre-defined benchmark-to-beat in an ensemble of split-sample tests. The method selection, whether spatial proximity-based or regression-based, has a greater impact on the regionalization than the specific details on how the method is applied. In particular, the descriptor selection plays a subsidiary role when at least a subset of selected descriptors contains relevant information.~~ Additionally, our research ~~has shown~~indicates that ~~regionalization~~ regionalization ~~causes~~ leads to spatially and temporally varying uncertainty ~~for~~ in ungauged regions. For example, ~~India and Indonesia are particularly affected by higher uncertaint~~regionalization highly affects southern South America, e.g., leading to high uncertainties in the flood simulation of the Río Deseado.~~y .~~The local impact of ~~regionalization~~ regionalization propagates through the water system, also affecting ~~in ungauged areas propagates through the water system~~global estimates, ~~e.g.,~~as evidenced by ~~one water balance component~~a ~~changed~~ spread of~~by approximately~~ ~~2400~~ 1,500 km³ yr^{-1} across an ensemble of five regionalization methods in simulated global runoff to the ocean~~on a global scale.~~, ~~which is in the range of inter-model differences.~~ This~~e~~ ~~magnitude of the impact of regionalization~~ discrepancy is even more pronounced when using a regionalization method deemed unsuitable for WaterGAP3, resulting in a spread of 4,208 km^3 yr^{-1}. This significant increase highlights the importance of carefully choosing regionalization

methods. Further research is needed to enhance the understanding of the methods' robustness on a global scale.~~depends on the variability in regionalized values and the region's sensitivity for the analysed component.~~

1. Introduction

Global hydrological models (GHMs) are developed and applied worldwide, e.g., to detect hotspots and examine patterns of climate change impacts on the terrestrial water cycle (e.g., Barbarossa et al., 2021; Boulange et al., 2021). Valid model results are a prerequisite to draw robust conclusions. For valid model~~l~~ing results, it is beneficial to adjust the parameter values to adapt the models to different basin processes (Gupta et al., 1998). This adaptation is usually modified and evaluated (in a loop) by comparing the simulated model output, often discharge, with the monitored data. However, this parameter adjustment for GHMs is challenging due to the lack of global monitoring data. Consequently, parameter adjustment for GHMs can be based not only on monitored data (i.e., calibration) but also on estimating parameter values for ungauged basins (i.e., ~~regionalization~~regionalization).

~~Regionalization~~ Regionalization defines ~~is~~ the estimation ~~of parameter values in a model~~of model parameters for ungauged basins (Oudin et al., 2008), usually based on information from gauged basins (Oudin et al., 2010). ~~Regionalization~~ Regionalization methods generally follow the same principle: basin characteristics (e.g., physiographic and/or climatic) are linked to hydrological characteristics and can thus be used to estimate parameter values. Various ~~regionalization~~ regionalization methods exist, and no overall preferred method has been found (Ayzel et al., 2017; Pool et al., 2021). In contrast, the optimal ~~regionalization~~ regionalization method may differ, for example, regarding available information (Pagliero et al., 2019) or model structures (Golian et al., 2021). Therefore, different methods should be tested to find an optimal ~~regionalization~~ regionalization method for a specific use case (e.g., Qi et al., 2020).

Evaluation is needed to assess different ~~regionalization~~ regionalization methods. The e~~E~~valuation of ~~is particularly challenging for regionalization~~ regionalization methods is particularly challenging ~~because~~ because they are usually applied when there is a lack of ~~monitoring data is missing~~monitoring data. Therefore, ~~regionalization~~ regionalization studies often treat gauged basins as ~~“~~"ungauged~~”~~" and perform leave-one-out cross-validation (e.g., Chaney et al., 2016) or split-sample tests (e.g., Beck et al., 2016; Nijssen et al., 2000; Yoshida et al., 2022). While at the mesoscale, this evaluation is already an integral part (e.g., McIntyre et al., 2005; Parajka et al., 2005; Oudin et al., 2008; Yang et al., 2020), this is sometimes not the case in global or continental studies (e.g., Müller Schmied et al., 2021; Widén-Nilsson et al., 2007). Another reasonable evaluation strategy is the concept of benchmark-to-beat (Schaefli & Gupta, 2007; Seibert, 2001). Applying a benchmark-to-beat supports a comprehensive evaluation of whether a new approach is functional, e.g., better than a straightforward and thus transparent method or better than a predecessor. To the authors' knowledge, such a benchmark-to-beat has never been used to evaluate innovations in ~~regionalization~~ regionalization at ~~the~~ a global ~~level~~scale.

In general, ~~regionalization~~ regionalization methods can be divided into two categories based on the parameter estimation strategy: (1) regression-based and (2) distance-based (He et al., 2011). Regression-based methods derive the relationship between basin characteristics and model parameters through fitted regression models. These mathematically defined relationships are further applied to estimate model parameters of ungauged basins (e.g., Kaspar, 2004; Müller Schmied et al., 2021). A significant drawback of regression-based ~~regionalization~~ regionalization is the difficulty of incorporating parameter interdependencies (Poissant et al., 2017), as ~~. R~~regression-based approaches often assume that the dependent variables, i.e., the model parameters, are not correlated (Wagener et al., 2004). Distance-based approaches transfer complete parameter sets from similar or nearby donor basins to ungauged basins (e.g., Beck et al., 2016; Nijssen et al., 2000; Widén-Nilsson et al., 2007). Using an ensemble of donor basins, e.g., by averaging the parameter values or model outputs, can improve the performance of such methods (e.g., Arsenault & Brissette, 2014). A significant disadvantage of such methods is the clustering problem of ungauged basins, i.e., the unequal distribution of gauging stations worldwide (Krabbenhoft et al., 2022). Thus, basins exist where distance-based approaches will use incomparable basins to transfer parameter values due to the lack of close basins.

Recent advances have implemented machine learning-based techniques in the context of ~~regionalization~~regionalization. For example, Chaney et al. (2016) used regression trees as an alternative to least squares regression to estimate parameter values in ungauged basins. Pagliero et al. (2019) explored supervised and unsupervised clustering methods to define the similarity of basins to transfer parameter sets. To the authors' knowledge, no study has compared several traditional ~~regionalization~~ regionalization methods with machine learning-based methods for a GHM on a global scale.

Some ~~regionalization~~ regionalization methods do not make a clear distinction between calibration and ~~regionalization~~regionalization. For example, Arheimer et al. (2020) applied a basin grouping beforehand. Then, they jointly calibrated the group members to define representative parameter sets. Subsequently, the representative parameter sets are transferred to other basins based on grouping rules. Another approach defines so-called transfer functions (Samaniego et al., 2010) and calibrates meta-parameters instead of the model parameter values (Beck et al., 2020; Feigl et al., 2022). These methods, where ~~regionalization~~ regionalization is part of the calibration process, often require a change in the calibration process itself, which is challenging for GHMs (Schweppe et al., 2022), for example, due to a lack of code flexibility (e.g., Cuntz et al., 2016).

This study proposes an improved ~~regionalization~~ regionalization method for the state-of-the-art GHM WaterGAP3 (Eisner, 2016). It compares traditional ~~regionalization~~ regionalization methods with machine learning-based methods and uses a ~~“~~benchmark-to-beat~~”~~ and an ensemble of split-sample tests to evaluate the applied methods. Further, global runoff simulations are compared to analyze the impact of regionalization methods. The overall research topic is evaluating and selecting ~~the most appropriate regionalization~~ regionalization methods for a GHM. Specifically, the study has two objectives. It aims

(1) to propose an improved ~~selection for the~~ regionali~~z~~zation method for~~of~~ WaterGAP3 and
(2) to evaluate the impact of ~~an improved regionalization~~ regionalization methods ~~against a benchmark-to-beat~~on global runoff simulations.

2. Data and Methods

2.1 The Model: WaterGAP3

The GHM WaterGAP3 simulates the terrestrial water cycle, including the main water storage components and a simple storage-based routing algorithm. It is a fully distributed model that operates on a five arcmin grid and simulates at a daily time step. A more detailed ~~model description can~~description of the model can be found in Eisner (2016).

In WaterGAP3, most model parameter values are set a priori, e.g., using look-up tables for albedo or rooting depth. Only one parameter, γ, is calibrated, which is part of the soil moisture storage in which runoff generation processes are present. The model equation for γ, which originates from the HBV-96 model (Lindström et al., 1997), is given in Eq. (1). Generally, higher values of γ lead to lower runoff volumes, while lower values of γ lead to higher runoff volumes. The~~is~~ model parameter is calibrated per basin within the range of 0.1 and 5. The objective function ~~for~~ of the calibration is to ~~minimize~~ minimize the deviation between the mean annual simulated and observed river discharge, i.e., the calibration aims to reduce the error in discharge volume. Given the monotonic relationship between the model's parameter and the optimization function, a simple search algorithm is applied: The parameter space is divided into rectangles, which are subsequently subdivided into smaller rectangles depending on the direction γ should be modified to achieve closer alignment with the optimization target. T~~Thus, as a result of t~~he calibration results in one ~~, each basin has a~~ calibrated γ value ~~(γ)~~ between 0.1 and 5 per basin. After the calibration, a correction is applied to account for high errors in the mass balance, e.g., due to inaccuracies in global meteorological forcing products. This correction ~~can only be applied in~~is only applicable on gauged basins. It is, therefore, neglected in this study.

$$R = P_t \cdot \left(\frac{S_s}{S_{s,max}}\right)^{\gamma} \tag{1}$$

where R is the daily runoff, P_t is the daily throughfall, S_s is the actual soil storage, $S_{s,max}$ is the maximal soil storage (given as a global map in Appendix A), and γ is the calibration parameter.

Traditionally, the regionalization process in WaterGAP3 is a simple multiple linear regression (MLR) approach to estimate the calibration parameter γ for ungauged basins (e.g., Döll et al., 2003; Kaspar, 2004). The drawback of MLR regarding parameter interaction can be neglected: As there is only one parameter to estimate, parameter interference does not exist. Instead, the approach offers the advantage of a lightweight, transparent application that can be quickly revised and adapted.

~~Traditionally, the regionalization process in WaterGAP3 is a simple multiple linear regression (MLR) approach to estimate the calibration parameter γ for ungauged basins (e.g., Döll et al., 2003; Kaspar, 2004). The drawback of MLR regarding parameter interaction can be neglected: As there is only one parameter to estimate, parameter interference does not exist. Instead, the approach offers the advantage of a lightweight, transparent application that can be quickly revised and adapted. We use the regionalization approach from WaterGAP2.2d as benchmark-to-beat as defined in Müller Schmied et al. (2021). WaterGAP2 has a model structure and calibration process that are very similar to WaterGAP3. The main difference between these models is that WaterGAP2.2d simulates at 0.5°spatial resolution. Thus, we expect the regionalization approach to be feasible for WaterGAP3.~~

2.2 Model Data

WaterGAP3 requires various input data, such as soil information, topography, or information on open freshwater bodies. This study uses the same input data as Kupzig et al. (2023). For meteorological forcing, we use the global data set EWEMBI (Lange, 2019). This data product includes daily global forcing data with a spatial resolution of 0.5 degrees (latitude and longitude) that covers a period from 1979 to 2016. Specifically, WaterGAP3 uses the following forcing information from the EWEMBI data set as input:

- daily mean temperature,
- daily precipitation,

- daily shortwave downward radiation, and
- daily longwave downward radiation.

The WaterGAP3 calibration requires observed monthly river discharge data. This discharge data is subsequently transformed into annual discharge sums ~~in the calibration procedure~~ and used as a benchmark in the calibration procedure. In this study, we used discharge data from 1,861 stations that were manually verified (Eisner, 2016). To get the best data available, we have updated all available station data with recent data from The Global Runoff Data Center (GRDC, 2020). All stations have at least five years of complete (monthly) station data between 1979 and 2016. For each station, a contribution area, i.e., a basin, is defined with the gridded flow-direction information obtained from WaterGAP3, ~~which is~~ based on the HydroSHEDS database (Lehner et al., 2008).

The 1,861 basins are calibrated using the above-described standard calibration approach for WaterGAP3. ~~After~~ Following the standard calibration procedure, some basins still have an insufficient model performance. In this context, we define a monthly Kling-Gupta-Efficiency (KGE) below 0.4 or more than 20 % bias in monthly flow as insufficient model performance. We underscore the importance of minimizing the error in discharge volume by defining it as an additional criterion corresponding to the optimization target during calibration. ~~, i.e., more than 20% bias in monthly discharge. These~~ Basins not fulfilling the defined conditions regarding bias and KGE ~~basins~~ are neglected in further analysis to avoid high parameter uncertainty due to errors in input data, model structure, or discharge data affecting the analysis~~—~~. Further, we have excluded all basins with less than 5000 km^2 (inter-) basin size ~~to~~ from the next upstream basin. We assume that this inter-basin size is large enough to assume a certain degree of interdependency between nested basins. In total, ~~1,236~~933 ~~basins~~ out of 1,861 basins are selected for ~~regionalization~~ regionalization (~~323~~ 626 are neglected due to insufficient ~~low~~ model performance, and 302 are neglected due to ~~insufficient~~ inadequate basin size).

Figure 1~~Figure 1~~a ~~shows a map of the~~depicts the worldwide calibrated basins, highlighting gauged and ungauged regions. Whereas~~,~~ ~~covering~~ most parts of North and South America are gauged,~~. However,~~ Africa and ~~Oceania~~ Australia remain largely ungauged. A cluster of gauged basins is ~~located~~ in Central Europe and in Eastern Asia. Gauged regions with ~~low~~ insufficient model performance are mainly ~~found~~ in the Mississippi River basin, Southern Africa, ~~and~~ Australia, and large parts of Brazil. These regions are known to be challenging for GHMs (e.g., cf. Fig. 8b in Stacke & Hagemann, 2021).

Figure 1~~Figure 1~~b shows the calibrated values for γ. It emerges that the calibrated values tend to be~~t~~ at the upper and lower bounds of the parameter space. This ~~mis~~behavio~~u~~r is already known (cf. Fig. 4b in Müller Schmied et al., 2021). A brief sensitivity analysis and discussion of the calibration parameter are included in Appendix B. The results of this analysis indicate that the clustering of the calibrated parameter value is not related to an inappropriate selection of the parameter bounds but instead to the absence or an insufficient representation of processes. Thus, the clustering of the calibrated values does not indicate an inadequate selection of the parameter bounds but ~~and~~ highlights the ~~need~~ necessity to ~~further develop~~improve the model structure and the calibration strategy for WaterGAP3~~, e.g., by implementing multivariate calibration~~. However, this study focuses solely on ~~analysing~~ analyzing and implementing ~~a new regionalization~~ regionalization methods. It does not aim to enhance the model structure or to change the calibration ~~approach~~ procedure of WaterGAP3. Future studies are needed to achieve the latter, as WaterGAP3 contains many hard-coded parameters or parameters defined by look-up tables that need to be analyzed to identify and adjust sensitive parameters more accurately during calibration. ~~To achieve the latter, future studies are needed to select sensitive parameters or advance the model structure to avoid structural errors that introduce high parameter uncertainty when applying multivariate calibration (Kupzig et al., 2023).~~ Initial steps in this direction have already been taken for WaterGAP2 in the form of a multivariate and multi-objective case study in the Mississippi River basin (Döll et al., 2024).

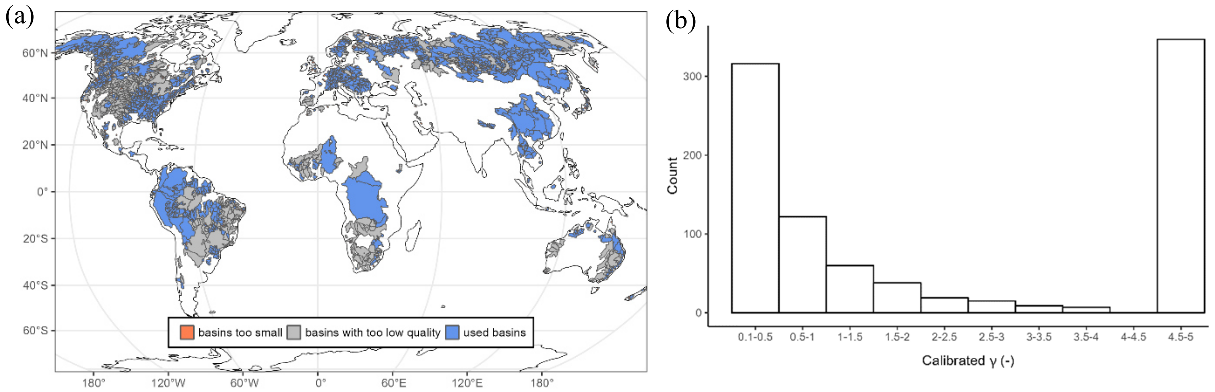


Figure 1: (a) Map of calibrated ~~Gauged~~ basins ~~calibrated beforehand~~, highlighting ~~highlighting~~ basins not used for ~~regionalization~~ regionalization due to ~~low~~ insufficient model performance or ~~too small~~inadequate basin size and (b) the histogram of the calibrated model parameter values of all used basins showing ~~heavy tails.~~a cluster of parameter values at the parameter bounds.

2.3 Basin Descriptors

This study uses basin descriptors as predictors to drive regression-based or distance-based ~~regionalization~~ regionalization approaches. These basin descriptors are based on ~~model~~ data used within the model simulation (as they are globally available). They~~and~~ are aggregated to basin values using a simple mean method to have the ~~exact~~ same spatial resolution as the calibrated model parameter.~~-~~ Thus, in the case of nested basins, the inter-basin area is used to define the basin descriptors. The selection of the predictors, i.e., basin descriptors that support the estimation of γ, is crucial for ~~regionalization~~ regionalization methods (Arsenault & Brissette, 2014). Typically, this selection aims to obtain the most information with the least number of predictors to (1) improve the model quality and (2) limit over-~~parametrization~~parametrization. In this study, we use 12 basin descriptors to develop ~~regionalization~~ regionalization methods; nine of these descriptors are physiographic, while the remaining three are climatic (see Table 1~~Table 1~~). Most descriptors are not correlated (see Appendix C~~A~~), i.e., we ~~avoid~~ minimize redundant information (Wagener et al., 2004).

A descriptor subset is selected based on correlation analysis between basin descriptors and calibrated γ value and entropy assessment. Pearson's correlation coefficient detects linear correlation, and Spearman's Rho and Kendall's Tau detect a non-linear correlation. Shannon entropy (Shannon, 1948) measures the information gain of the predictors explaining the calibrated γ value. The higher the information gain, the more valuable the basin descriptor is for explaining the variation in the calibrated γ value. The analysis directly evaluates the relationship between the calibrated parameter and the basin descriptors, as WaterGAP3 uses only one calibration parameter with a clear global optimum within the parameter space. An alternative would be to use flow characteristics to define the basis for regionalization (e.g., Pagliero et al., 2019). We decided to use the calibrated parameter instead of flow characteristics as it does not need any further assumption on which flow characteristics determine the model's parameter. ~~The predictor selection is based on correlation analysis and entropy assessment. Pearson's correlation coefficient detects linear correlation, and Spearman's Rho and Kendall's Tau detect a non-linear correlation between basin~~

~~descriptors and calibrated γ values. Shannon entropy (Shannon, 1948) measures the information gain of the predictors explaining the calibrated γ value. The higher the information gain, the more valuable the basin descriptor is for explaining the variation in the calibrated γ value.~~

Statistical information of the evaluated basin descriptors and the corresponding ~~The c~~correlation coefficients and ~~the corresponding~~ information gain are listed in Table 1 ~~Table 1~~. The basin descriptors demonstrate a considerable degree of variability, e.g., the basin size ranges from 5000 km^2 to 3,112,480 km^2 with a median of 13,796 km^2. The mean temperature varies from -19 °C to 29 °C, and the sum of precipitation ranges from 213 mm to 5,716 mm. Although there is a high degree of variability in the analyzed basin descriptors, ~~All~~ the basin descriptors ~~have~~ exhibit~~a~~ low correlation coefficients with the calibrated values. ~~, e.g.~~For example, the permafrost coverage shows the ~~highest~~ strongest Pearson correlation of ~~is~~ -0.37 (and -0.50 for Spearman's Rho)~~6~~. The information gain indicates the same results as the correlation analysis, i.e., the information gain is generally relatively low, and descriptors with a higher correlation tend to have a higher information gain.~~The information gain shows the same result for the predictors, i.e., descriptors with a higher correlation tend to have a higher information gain. Nevertheless, the information gain is relatively low~~ For example, the mean temperature exhibits the maximal information gain~~, with a maximum~~ of 17.6 ~~4.4~~% and has the second-highest correlation coefficient with a Pearson correlation of 0.34~~of the information explained by the temperature descriptor~~.

Table 1: Basin descriptors: statistical information, correlation, and entropy assessment. Selected physiographic and climatic basin descriptors are written in bold.

	Basin	Attribute Information				Entropy & Correlation			
	Descriptor	Min	Max	Mean	Median	IG (%)[1]	Pearson	Spearman	Kendall
physiographic	Soil Storage (mm)	12.405	610.469	220.805	195.778	13.07	-0.21	-0.15	-0.11
	Open Water Bodies (%)	0.000	63.960	5.521	1.812	5.65	-0.01	-0.08	-0.05
	Wetlands (%)	0.000	63.466	4.164	0.547	5.01	-0.02	-0.13	-0.09
	Size (km^2)	5000	3,112,480	37,572	13,796	1.42	-0.04	-0.04	-0.03
	Slope Class (-)	10.057	67.756	38.668	38.364	16.60	-0.31	-0.37	-0.27
	Altitude (m.a.s.l.)	30.239	4765.166	591.024	394.870	9.30	-0.18	-0.28	-0.20
	Sealed Area (%)	0.000	12.3	0.6	0.1	4.49	0.22	0.38	0.29
	Forest (%)	0.000	100.000	35.340	24.002	13.82	-0.25	-0.18	-0.14
	Permafrost & Glacier (%)	0.000	95.000	16.662	0.000	13.12	-0.37	-0.50	-0.40
climate	**Mean Temperature(°C)**	-18.848	28.823	7.720	7.707	17.56	0.34	0.41	0.30
	Yearly Precipitation (mm)	213.6	5,716.3	996.5	779.5	9.23	0.02	0.21	0.14
	Yearly Shortwave Downward Radiation (Wm^{-2})	1,050.6	3,043.2	1,857.9	1,759.7	15.79	0.31	0.33	0.24

[1] Information gain is given in percentage of total information content in γ after Shannon (1948)

In contrast to the findings of Wagener and Wheater (2006), the correlation coefficients between the basin descriptors and the calibrated values are relatively low, indicating a weak relationship. One potential explanation for this discrepancy is that Wagener and Wheater (2006) used a smaller number of basins in southeast England, with limited versatility (e.g., regarding climate and seasonality) compared to the 933 worldwide basins used in this study. Studies using a large number of basins likely tend to find a lower correlation between catchment attributes and model parameters (Merz et al., 2004). Moreover, the clustered calibrated γ values at the bounds of the valid parameter space may disturb the results of this analysis. ~~A possible reason for the low correlation and information gain is that the γ values are tailored within the calibration's valid parameter bounds (i.e., 0.1 and 5), resulting in heavy tails of the calibrated γ distribution. Thus, we expect the correlation to be higher, with calibrated γ reaching values higher than 5. In addition,~~As the calibrated value masks the effect of multiple sources of errors, such as uncertainty in the input data, model structure, or varying hydrological processes, finding a meaningful relationship between catchment characteristics and calibrated values is challenging.

Because the basis for the descriptor selection seems uncertain, given the low correlation and the named constraints, we additionally run the regionalization methods with all descriptors to evaluate the descriptor selection. Further on, to ascertain the advantage of integrating climatic descriptors, we run the regionalization methods using either physiographic or climatic descriptors. ~~Thus, there might be more complex relationships between the descriptors and the calibrated parameter, which are only partially captured by this analysis. Nevertheless, the results of this analysis indicate descriptors that may be more useful than others in defining a regionalization method.~~ In total, ~~W~~we used ~~implement regionalization methods using~~ four groups of basin descriptors to implement the regionalization methods ~~by selecting basin descriptors with the highest correlation coefficients and information gain~~:

- ~~“~~"cl~~”:~~ ": ~~two correlated~~all three climatic descriptors,~~ (mean temperature, annual shortwave radiation),~~
- ~~“~~"p~~”:~~ ": ~~three correlated~~all nine physiographic descriptors ~~(slope class, forest %, permafrost %)~~,
- "p+cl": all 12 descriptors, and
- ~~“~~"~~p+cl~~subset~~”:~~ ": two correlated climatic descriptors (mean temperature, annual shortwave radiation) & three correlated physiographic descriptors (slope class, forest %, permafrost %).~~, and~~
- ~~“all”: all 12 descriptors (as a control group to examine the effect of using correlated descriptors).~~

~~Table 1: Basin descriptors used in the regionalization methods: statistical information, correlation, and entropy assessment. Selected physiographic and climatic basin descriptors are shaded in grey.~~

	~~Basin Descriptor~~	~~Attribute Information~~				~~Entropy & Correlation~~			
		~~Min~~	~~Max~~	~~Mean~~	~~Median~~	~~IG (%)~~	~~Pearson~~	~~Spearman~~	~~Kendall~~
~~physiographic~~	~~Soil Storage (mm)~~	~~8.994~~	~~677.950~~	~~219.071~~	~~192.006~~	~~10.19~~	~~-0.20~~	~~-0.16~~	~~-0.12~~
	~~Open Water Bodies (%)~~	~~0.000~~	~~77.125~~	~~7.979~~	~~2.376~~	~~5.22~~	~~0.01~~	~~-0.05~~	~~-0.03~~
	~~Wetlands (%)~~	~~0.000~~	~~73.181~~	~~6.134~~	~~0.721~~	~~4.60~~	~~0.02~~	~~-0.07~~	~~-0.05~~
	~~Size (km^2)~~	~~5000~~	~~3112480~~	~~36811~~	~~13850~~	~~1.08~~	~~-0.03~~	~~-0.01~~	~~-0.01~~
	~~Slope Class (-)~~	~~10.057~~	~~67.756~~	~~37.739~~	~~36.986~~	~~14.22~~	~~-0.27~~	~~-0.31~~	~~-0.23~~
	~~Altitude (m.a.s.l.)~~	~~22.324~~	~~4765.166~~	~~630.826~~	~~412.414~~	~~7.29~~	~~-0.11~~	~~-0.19~~	~~-0.14~~
	~~Sealed Area (%)~~	~~0.000~~	~~12.3~~	~~0.5~~	~~0~~	~~3.25~~	~~0.18~~	~~0.34~~	~~0.25~~
	~~Forest (%)~~	~~0.000~~	~~100.000~~	~~32.037~~	~~18.245~~	~~11.50~~	~~-0.27~~	~~-0.21~~	~~-0.16~~
	~~Permafrost & Glacier (%)~~	~~0.000~~	~~95.000~~	~~15.316~~	~~0.000~~	~~10.96~~	~~-0.36~~	~~-0.47~~	~~-0.37~~
~~climate~~	~~Mean Temperature(°C)~~	~~-18.848~~	~~28.998~~	~~7.769~~	~~6.562~~	~~14.36~~	~~0.34~~	~~0.39~~	~~0.29~~
	~~Yearly Precipitation (mm)~~	~~73.1~~	~~5716.3~~	~~950.6~~	~~743.5~~	~~7.95~~	~~0,01~~	~~0.18~~	~~0.13~~
	~~Yearly Shortwave Downward Radiation (Wm^{-2})~~	~~1050.6~~	~~33098.4~~	~~1887.5~~	~~1777.2~~	~~13.05~~	~~0.33~~	~~0.34~~	~~0.25~~

2.4 ~~Regionalization~~ Regionalization Methods

In our study, we test several traditional and machine learning-based ~~regionalization~~ regionalization methods against each other and a defined benchmark-to-beat to find ~~the most~~ suitable ~~regionalization~~ regionalization methods for WaterGAP3. At the global scale, ~~regionalization~~ regionalization is particularly challenging due to (1) the lack of high-quality data, (2) the diversity of dominant hydrological processes in basins, and (3) the high computational demands of the models. Therefore, a robust regionalization method that applies to a wide variety of basins and is not computationally demanding should be selected for a global application. ~~Therefore, a regionalization method that is robust, applicable to a wide variety of basins, and not computationally demanding should be chosen.~~

We test three common traditional approaches and two machine learning-based approaches using the concepts of spatial proximity, physical similarity, and regression-based methods. As WaterGAP3's model calibration is very rigid and has only one parameter, it is not feasible to implement and test regionalization methods that incorporate regionalization into the calibration process, such as transfer functions. In addition, we avoid high computational demands as all evaluated methods are applicable after the calibration, i.e., without running the model.~~We test three common traditional approaches: spatial proximity, physical similarity, and regression-based methods, as well as two machine learning-based approaches. These machine learning-based approaches are alternatives to traditional physical similarity and regression-based methods. As the model calibration of WaterGAP3 is very rigid and has only one parameter, it is not feasible to implement and test regionalization methods that incorporate regionalization into the calibration process, such as transfer functions. In addition, we avoid high computational demands as all methods can be applied after the calibration, i.e., without running the model.~~

As the calibration of WaterGAP3 results in a parameter distribution with a cluster of parameter values at the parameter bounds, we implement a so-called "tuning" to introduce information about the parameter space into regionalization. In detail, we apply a simple threshold-based approach to shift the regionalized parameter values to the extremes, i.e., $\gamma_{est} < \gamma_1 \rightarrow \gamma_{reg} = 0.1$ and $\gamma_{est} > \gamma_2 \rightarrow \gamma_{reg} = 5.0$. The thresholds γ_1 and γ_2 are defined by applying the k-means algorithm with three centers to the calibrated parameter values. This clustering results in three clusters: one for low, one for medium, and one for high γ values. Subsequently, γ_1 refers to the highest γ value of the low cluster and γ_2 refers to the lowest γ value of a high cluster.

~~To~~ To evaluate~~evaluate~~ the ~~regionalization~~ regionalization methods, we implement an ensemble of split-sample tests. Specifically, we randomly split the basins into 50 % gauged (for training) and 50 % pseudo-ungauged (for testing)~~basins~~. The~~is~~ split has a relatively high percentage of pseudo-ungauged basins, accounting for many missing gauges worldwide. We fit the methods and apply them to the training and testing data sets. The split-sample test is repeated 100 times ~~with~~ by randomly ~~selected~~ splitting the basins ~~basins for training and testing~~ to account for sampling effects.

As there is only one calibration parameter, γ, this parameter has a global optimum per basin. Consequently, the quality of training and testing is directly assessed by the deviation between the ~~predicted~~ regionalized and the calibrated value for γ. The closer the regionalized values are to the calibrated ones, the more accurate the prediction. We assess the prediction accuracy by the logarithmic version of the mean absolute error (logMAE) to account for the decreasing sensitivity of γ for higher values (see Appendix B). ~~Thus, the mean absolute error (MAE), an easy-to-interpret measure, is used to evaluate the prediction accuracy.~~ The lower the logMAE, the better the prediction; ~~an~~a ~~MAE~~ zero value~~of zero~~ in logMAE expresses no error. ~~In our case, an MAE of 2.1 corresponds to the error when using the mean calibrated γ value as the predicted value.~~ The ~~regionalization~~ regionalization method is robust if the prediction accuracy is similar in training and testing. A generally good performance, i.e., small logMAE values, indicates that the ~~regionalization~~ regionalization method suits WaterGAP3. The comparison of γ values enables applying a wide range of regionalization methods and sets of descriptors, as no computationally intensive model simulation is required. However, it assumes that deviations in γ lead, in turn, to deviations in discharge, which is only partially true because of varying parameter sensitivity in basins (e.g., Kupzig et al., 2023). To validate that the logMAE is a sufficient approximator for the regionalization performance in WaterGAP3, we use one representative split-sample from the ensemble to compare the accuracies in simulated discharge for different regionalization methods.

Regression-based methods

The traditionally used regionalization approach of WaterGAP3 is a regression-based MLR. As the benchmark-to-beat, we use the regionalization approach from WaterGAP2.2d defined in Müller Schmied et al. (2021). We consider it a suitable benchmark-to-beat given that WaterGAP2 has a model structure and calibration process that is very similar to WaterGAP3. The main difference between these models is that WaterGAP2 simulates at 0.5°spatial resolution. The benchmark-to-beat consists of "a multiple linear regression approach that relates the natural logarithm of *γ* to basin descriptors (mean annual temperature, mean available soil water capacity, fraction of local and global lakes and wetlands, mean basin land surface slope, fraction of permanent snow and ice, aquifer-related groundwater recharge factor)". (Müller Schmied et al., 2021) We fit this regression model to our data and define the quality of this approach as the benchmark-to-beat. Moreover, we test an independent MLR approach without using the logarithmical scaling of *γ* and using the above-defined sets of basin descriptors. For MLR and the benchmark-to-beat, we use the lm() function of the R package stats (R Core Team, 2020). After applying the regression model, we adjust the estimated parameter values to ensure that the estimated values range between 0.1 and 5.

~~For the traditional regression-based methods, we use the lm() function of the R package stats (R Core Team, 2020) to implement an MLR. After applying the regression model, we adjust the estimated parameter values to ensure that the estimated values range between 0.1 and 5. As the calibration of WaterGAP3 results in a parameter distribution with heavy tails, we implement a so-called "tuning approach" to introduce this information into regionalization. In detail, we apply a simple threshold-based approach to adjust the regionalized parameter values to the extremes, i.e.,~~ ~~$\gamma_{est} < \gamma_1 \rightarrow \gamma_{reg} = 0.1$~~ ~~and~~ ~~$\gamma_{est} > \gamma_2 \rightarrow \gamma_{reg} = 5.0$~~~~. A simple clustering, i.e., the k-means algorithm with three centres, defines these thresholds.~~

Furthermore, a machine learning-based method, ~~namely~~ random forest (RF), is tested for ~~regionalization~~regionalization as an alternative to MLR. Here, we implement the random forest algorithm with the randomForest() function from the R package randomForest (Liam & Wiener, 2002), which is based on Breimann (2001). The algorithm uses an ensemble of decision trees, making the decision human-like. It is relatively robust because it incorporates random effects into the training process. To implement this randomness, we define th~~at the algorithm~~e algorithm as one that can choose between two randomly selected predictors at each node~~. We use an~~, using an ensemble of 200 trees.~~, the same combinations of predictors and the same tuning as for MLR.~~

~~The benchmark-to-beat defined in Müller Schmied et al. (2021) also uses an MLR approach. This MLR approach relates the natural logarithm of γ to the following basin descriptors: mean temperature, mean available soil water capacity, fraction of open freshwater bodies, mean slope, mean fraction of permafrost coverage and an aquifer-related groundwater recharge factor. Thus, the main differences between the benchmark-to-beat and our defined MLR-based approach are the natural logarithm, our proposed tuning procedure for the method itself, and using the aquifer-related groundwater recharge factor as a basin descriptor.~~

Physical Similarity

~~For~~ As ~~a~~ the traditional physical similarity approach, we use Similarity Indices (in the following named with SI), applying~~. We~~ ~~use~~ the methodology proposed by Beck et al. (2016). The SI (see Eq. (2)) are derived using the defined basin descriptors sets,~~mentioned above,~~ and the parameter of the most similar basin is transferred to the pseudo-ungauged basin. Additionally, we use an ensemble of basins to control whether an ensemble-based approach leads to more robust results. The optimal number of donor basins may vary between research regions and hydrological models (Guo et al., 2020). Here, we use ten donor catchments (noted with ~~“~~"ensemble~~10”), ~~") ~~which is ~~based on Beck et al. (2016) and McIntyre et al. (2005~~6~~). Further, we apply a simple mean method for the ensemble-based prediction to aggregate the ensemble of γ values into one predicted parameter value.

$$S_{i,j} = \sum_{p=1}^{n} \frac{|Z_{p,i} - Z_{p,j}|}{IQR_p} \tag{2}$$

where $S_{i,j}$ is the Similarity Index between basin i and basin j, $Z_{p,j}$ is the basin descriptor p for basin j, IQR_p is the interquartile range for basin descriptor p among all (gauged) basins, and n is the number of all basin descriptors used.

As an alternative ~~a~~ machine learning-based approach, we apply a simple k-means algorithm. We selected the k-means algorithm because it is one of the most widely used clustering algorithms (Tongal & Sivakumar, 2017). It is easy to understand and use. The algorithm kmeans() is implemented in the R base package stats. It aims to ~~maximize~~ maximize variation between groups and ~~minimize~~ minimize variation within groups. The number of clusters to use is determined by multiple indices calculated with the R package NbClust (Charrad et al., 2014). For all 933 basins and the defined sets of basin descriptors, most indices defined three as the optimal number of clusters. Accordingly, w~~W~~e use three clusters to generate the groups of basins. As different scales of the predictor values can affect the clustering, a rescaling with min-max-~~normalization~~ normalization (see Eq. (3)) is performed on the training set and applied to the testing set. After the grouping, the mean γ value is assigned as a representative calibrated value to the corresponding basin group. To estimate the corresponding group for a pseudo-ungauged basin, the knn algorithm is used, and the representative γ value of the group is assigned to the pseudo-ungauged basin. This algorithm is implemented by the knn() function of the R package class (Venables & Ripley, 2002). Since ~~this method~~the k-means method is less flexible than SI, we implement a highly flexible version, using the knn algorithm directly to define the donor basin most similar to each ungauged basin. ~~of k-means with 162 groups, where each ungauged basin is sorted into a very small basin group.~~ Using ~~this highly flexible version~~the knn algorithm directly ~~of k-means~~, we test ~~whether the potential differences between SI and k-means are based on the degree of flexibility~~how beneficial it is to create groups of similar basins using the kmeans algorithm and regionalize the parameter with a representative mean value.

$$Z'_{p,j} = \frac{Z_{p,j} - \underset{j\to m}{min}(Z_{p,j})}{\underset{j\to m}{max}(Z_{p,j}) - \underset{j\to m}{min}(Z_{p,j})} \tag{3}$$

where $Z'_{p,j}$ is the ~~normalise~~n~~d~~ormalized basin descriptor p for basin j, $Z_{p,j}$ is the basin descriptor p for the basin j, m is the number of (gauged) basins.

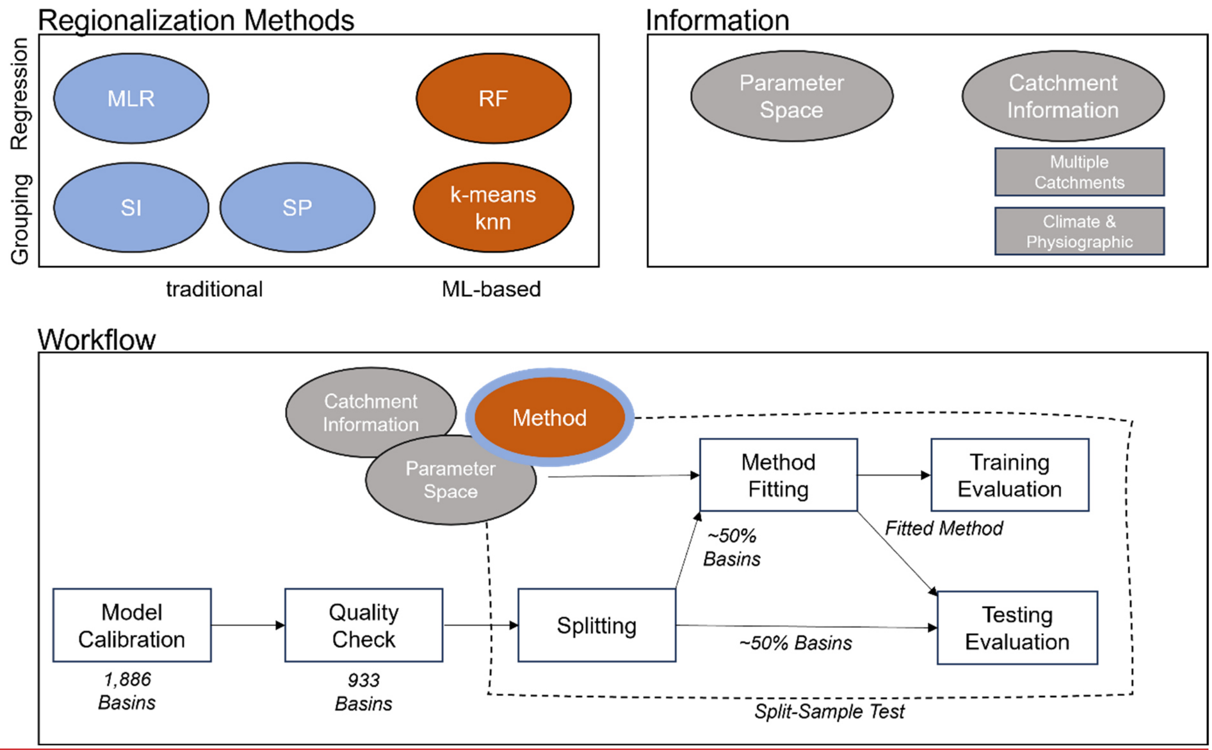
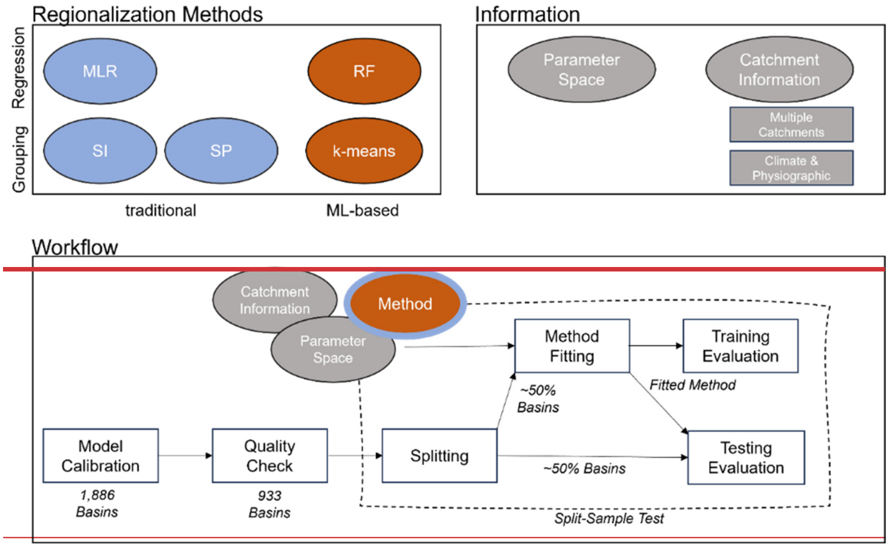


Figure 2: Experimental setup of the study: regionalization methods, used modifications and information, and the general workflow (MLR: Multiple Linear Regression, SI: Similarity Indices, SP: Spatial Proximity, RF: RandomForest).



~~**Figure 2: Experimental setup of the study: regionalization methods, used modifications and information and the general workflow (MLR: Multiple Linear Regression, SI: Similarity Indices, SP: Spatial Proximity, RF: RandomForest).**~~

Spatial Proximity

The spatial proximity approach is one of the easiest to ~~regionalize~~ regionalize parameter values. However, it is also often ~~criticized~~ criticized that nearby basins do not necessarily have the same hydrological behavio~~u~~r (Wagener et al., 2004). Furthermore, its performance depends on the density of the network of gauged basins (Lebecherel et al., 2016). The dependency on network density is particularly challenging for global applications where large parts of the world are ungauged (e.g., northern Africa). Nevertheless, the approach has been successfully applied in other studies (e.g., Oudin et al., 2008; Qi et al., 2020), even globally (Widén-Nilsson et al., 2007).

Here, we take the distance between the centroids of the basins as the~~a~~ reference for the spatial distance between basins, as done by others (Oudin et al., 2008). We use the abbreviation SP in the text below to refer to the spatial proximity approach. Figure 2 ~~Figure 2~~ provides an overview of the applied ~~regionalization~~ regionalization methods and information used for the experimental setup.

~~3. Results and Discussion~~

3. Results and Discussion

3.1 Evaluating the effect of ~~Traditional Methods~~tuning

First, the impact of the tuning approach on the regionalization approaches is evaluated. Therefore, Fig. 3 depicts the differences in logMAE between the standard and tuned approaches in testing, i.e., using the pseudo-ungauged basins. A positive difference in logMAE indicates an increase in accuracy, whereas a negative difference indicates a decrease in accuracy due to the tuning.

Using the tuning thresholds of about 1.1 and 3.4 for γ_1 and γ_2, respectively, enhances the predictive accuracy for kmeans, MLR, RF, and the ensemble approach of SI. The most remarkable improvement for kmeans, RF, and SI ensemble is achieved when all physiographic descriptors are used as input (mean improvement of 0.077, 0.058, and 0.071, respectively). MLR shows the most significant improvement when using all available descriptors (mean improvement of 0.038). In contrast, the tuning decreases the performance for knn, SI, and SP, with a mean degradation between -0.02 and -0.05. Unlike the enhanced regionalization techniques, these methods transfer single-basin information to ungauged regions. Thus, the tuning disturbs the use of single-basin information yet simultaneously enhances the performance of methods that transfer multi-basin information. The disturbance or improvement is probably related to the capability of the methods representing the clustering of parameter values at the extremes: Whereas the multi-basin information transfer implies a smoothing and thus suffers from a lack of representing the extremes, the single-basin information transfer exhibits no such a smoothing.

The exception from the above-defined rule is the benchmark-to-beat approach. The benchmark-to-beat is the only approach that uses logarithmic scaled γ values when fitting the model. This logarithmic transformation leads to an increase in estimating small values. Thus, when the benchmark-to-beat is tuned, more basins with higher calibrated γ values receive low estimates. The tuning intensifies this effect, leading to a decrease in the accuracy of the logMAE from the standard to the tuned version. Thus, for models using logarithmical transformed γ values, the defined thresholds for the tuning are not appropriate.

Applying knowledge of the optimal parameter space enhances the quality of regionalization for methods transferring multi-basin information in case the tuning thresholds are appropriate. This positive effect is not surprising, as incorporating a priori information about parameter distribution strengthens parameter estimation (e.g., described in Tang et al. (2016) using the Bayes Theorem). However, for single-basin transfer, which already represents the parameter space well, i.e., the clustering of γ at the extremes, the tuning disturbs the performance. This indicates that such tuning needs to be cautiously introduced as there is the risk of decreasing the accuracy of regionalization.

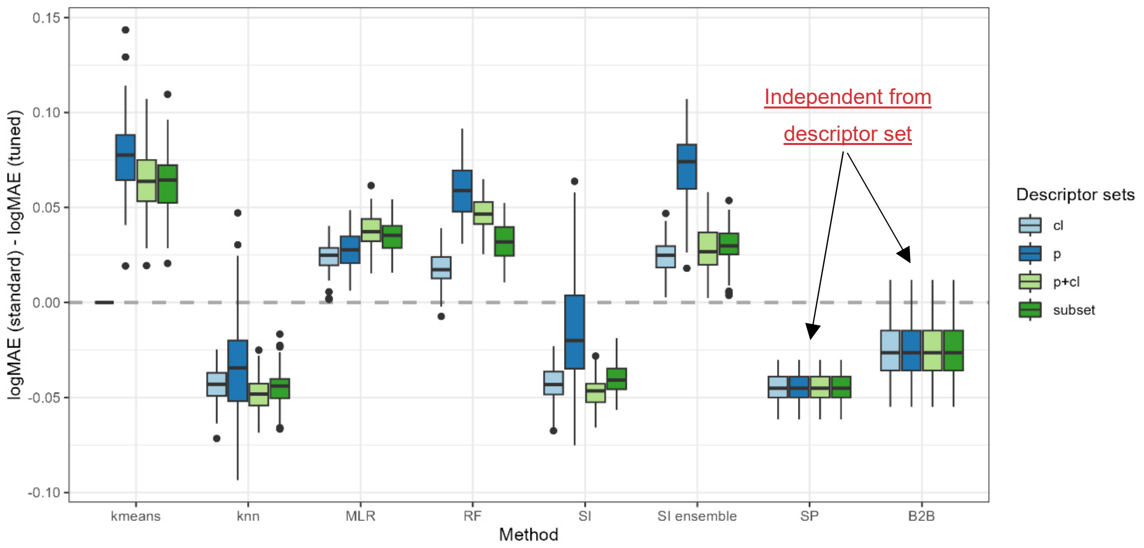


Figure 3: Changes in performance between standard and tuned versions for all applied regionalization approaches. Positive values indicate an improvement related to the tuning.

3.2 Evaluating descriptor subsets & algorithm selection

Different descriptor sets yield different performances in regionalizing γ. Table 2 shows the median of all logMAE values for the testing. For a complete overview of the results of the split-sample test ensemble, see Appendix D. Evaluating Table 2 reveals that the selected subset or all descriptors consistently yield the best performance across all regionalization methods. In both variants of the ensemble approach of SI, the tuned version of the no-ensemble approach of SI, and the standard version of RF, the selected subset yields the best results. For all other methods, using all descriptors yields the best results. Hence, all methods perform best when combining climatic and physiographic descriptors. This benefit of using climatic and physiographic descriptors is consistent with others that often apply a combination of climatic and physiographic descriptors, achieving optimal regionalization results (e.g., Oudin et al., 2008; Reichl et al., 2009).

The machine learning-based approaches seem to benefit most when using more information displaying an improvement for all methods (knn, kmeans, and RF) and both variants (standard and tuned) ranging from "cl", "p", "subset" to "p+cl". This is not surprising as machine learning is developed to deal with big data sets. The traditional methods MLR and SI do not exhibit such a distinct pattern. The (weakly) correlated subset of climatic and physiographic descriptors yields the best results for SI. As utilizing all descriptors decreases the performance slightly, the results indicate that uncorrelated descriptors may disturb the performance of this approach. For MLR, the meaning of physiographic information is highest, resulting in the best ("p+cl") and second best ("p") results. The disparate performance of the regionalization methods when using different descriptor sets indicates that different methods use descriptor sets with varying efficiency. It also emphasizes that the selection of descriptors impacts the regionalization method's results, as noted by others (Arsenault & Brissette, 2014). Consequently, the above-performed analysis defining a descriptor subset lacks universal validity as methods exist where the defined subset is outperformed. Instead, the validity of this approach is most closely aligned with the SI approaches.

Although the algorithms kmeans and knn are similar, they yield considerably different performances in Table 2. As knn shows a logMAE of 0.432 at best, the kmeans algorithm performs poorly, resulting in the best logMAE of 0.472. This indicates that applying the kmeans clustering algorithm to transfer averaged parameters is inappropriate for WaterGAP3. This may be attributed to the reduced flexibility of the approach, which entails estimating only three γ values due to the optimal, though limited, number of centers. The ensemble SI approach consistently outperforms the no-ensemble SI approach in almost all variants. The positive effect of an ensemble approach for SI has already been noted (Oudin et al., 2008). Therefore, it is recommended that the number of donor basins derived from the literature be adopted in future applications to be optimal for WaterGAP3, likely resulting in higher performance.

Only a few regionalization methods outperform the benchmark-to-beat. The best descriptor sets of tuned MLR, RF, and SI ensemble approach have a logMAE of 0.427, 0.403, and 0.409, respectively. The standard version of knn ("p+cl") and SP yield 0.432 and 0.454 in logMAE, respectively. Additionally, two variants of the standard SI approaches outperform the benchmark-to-beat yet exhibit inferior results compared to the selected tuned approach. All other regionalization methods show higher logMAE values than the benchmark-to-beat. These methods are considered insufficient in terms of performance to regionalize γ in WaterGAP3. As the benchmark-to-beat outperforms all kmeans approach variants, it is deemed unsuitable for regionalizing γ for WaterGAP3 and, therefore, excluded from further analysis.

Table 2: Median logMAE of 100 split-samples for pseudo-ungauged basins, i.e., in testing, for all regionalization methods applying four sets of descriptors for a) the standard version and b) the tuned version. The bold numbers indicate a better performance than the benchmark-to-beat. Thicker edges mark best-performing variants, which are chosen for further analysis. Grey-shaded cells indicate worst-performing variants, which were taken to validate the assumption that lower logMAE values result in lower KGE values.

(a) test (median)	MLR	RF	SI		kmeans	knn	SP	B2B
			no ens.	ensemble				
cl	0.552	0.483	0.496	0.483	0.619	0.501	**0.454**	0.461
p	0.479	0.465	0.487	0.480	0.551	0.477		
p+cl	0.464	0.464	**0.454**	0.462	0.534	**0.432**		
subset	0.488	0.488	0.461	**0.439**	0.539	0.467		

(a)(b) test* (median)	MLR	RF	SI		kmeans	knn	SP	B2B
			no ens.	ensemble				
cl	0.529	**0.467**	0.537	**0.459**	0.619	0.546	0.502	0.488
p	**0.441**	**0.416**	0.532	**0.455**	0.515	0.521		
p+cl	**0.427**	**0.403**	0.503	**0.435**	0.472	0.480		
subset	**0.453**	**0.408**	0.501	**0.409**	0.477	0.509		

The well-performing SP on a global scale is surprising as the distances between basins are potentially long, and hydrological processes may strongly vary. It is probably beneficial for the SP approach that γ comprises all kinds of errors, e.g., spatially localized errors in global forcing products (e.g., Beck et al., 2017 reported errors for arid regions in the precipitation product) or inaccurately represented processes for larger regions. Thus, the estimation of γ might be appropriate, but not because of the same hydrological behavior but due to the same kind of errors.

The RF approach is outstanding, as it shows a massive loss in performance from training to testing (see Appendix D). In detail, the logMAE in testing is about twice the logMAE in training. In comparison, other methods show results from 95.6 % to 101.4 %. This performance loss indicates that RF is not a robust regionalization method for WaterGAP3. Other studies that reported the good performance of RF for regionalization have not investigated the stability of the performance from training to testing (Golian et al., 2021; Wu et al., 2023). Likely, the mathematical problem of predicting the calibrated parameter for WaterGAP3, with all its challenges (e.g., tailored parameter space, clustered calibrated parameter, and incorporation of many sources of errors), cannot be adequately solved by RF. Thus, although RF is known to be especially robust among other machine learning-based techniques, it shows symptoms of over-parameterization. This indicates that the algorithm is too flexible and adjusts to noise in the data, missing the underlying systematic. This lack of robustness is particularly disadvantageous since, for WaterGAP3, regionalization is applied globally, requiring regionalizing large parts of the world. In consequence, the RF approach is left out from further analysis and defined as not suitable to regionalize γ for WaterGAP3.

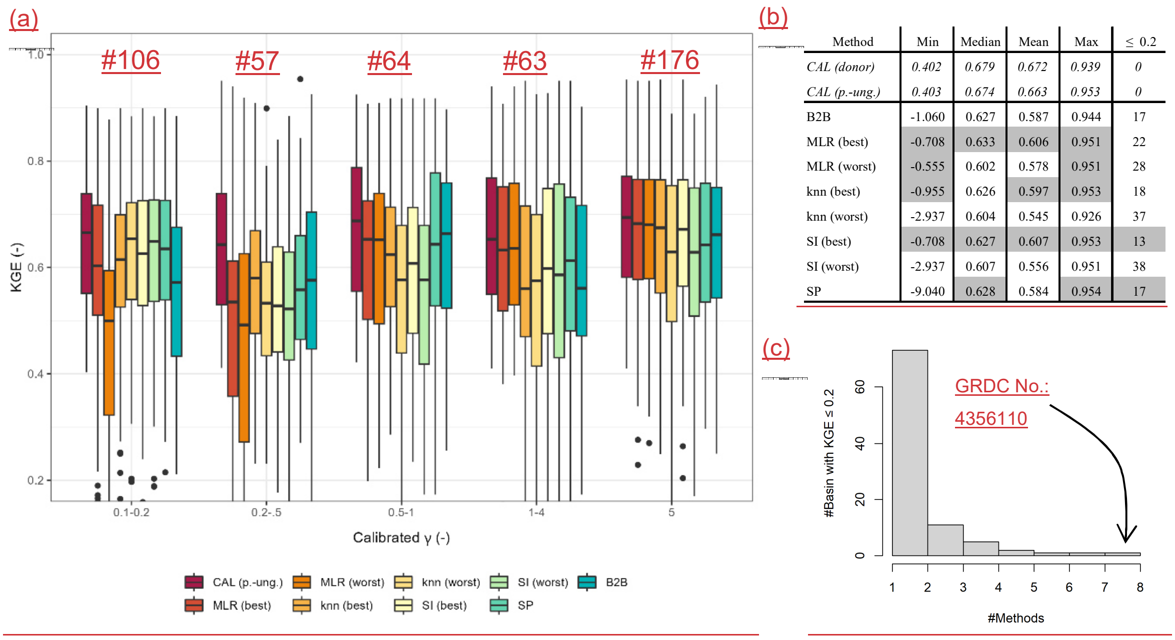
3.3 Performance of selected algorithm in pseudo-ungauged basins

To avoid the high risk of sampling effect when applying the split-sample test, we conduct an ensemble of 100 split-sample tests analyzing the median of logMAE between regionalized and calibrated values as an indicator for performance. Directly using the differences in regionalized and calibrated values is only meaningful when the calibrated value represents the global optimum. As this is often not the case, e.g., due to equifinality, the performance of regionalization methods is usually assessed by the accuracy of simulated discharge (e.g., Samaniego et al., 2010; Arsenault & Brissette, 2014). Because WaterGAP3 requires computationally intensive simulations, running WaterGAP3 for all 100 split-sample tests for the selected methods is not feasible. Therefore, we select a single representative split-sample to assess the quality of representing the discharge in the pseudo-ungauged basins using regionalized γ values. The representative split-sample leads to comparable logMAE values to the corresponding median of the ensemble for all regionalization methods. For the evaluation, WaterGAP3 was run for the same period used in calibration (from 1979 to 2016), with the first year simulated ten times to allow for model warm-up. Using this period ensures the availability of sufficient data for the evaluation (see Chapter 2.2). Furthermore, the differences between the monthly simulated and observed discharge are assessed using the KGE.

To evaluate the KGE, we select the best-performing methods that outperform the benchmark-to-beat: tuned MLR "p+cl", knn "p+cl", tuned SI ensemble "subset", and SP (see Table 2). For the sake of simplicity, we further mark them with "(best)". Additionally, we select three poorly performing variants to validate the assumption that methods resulting in higher logMAE values tend to result in lower KGE values, i.e., lower accuracy of simulated discharge. These methods are tuned SI "cl" (logMAE: 0.537), tuned knn "cl" (logMAE: 0.546), and MLR "cl" (logMAE: 0.552). Further, we denote these methods with "worst". Applying the selected methods and the benchmark-to-beat method results in eight estimates of γ for the pseudo-ungauged basins, whose performance is further evaluated in terms of simulated discharge accuracy.

Figure 4a shows the resulting KGE values for the evaluated regionalization methods and the calibrated version as grouped boxplots for different ranges of calibrated γ. The methods show different performances for different γ ranges, indicating their strengths and weaknesses. For the smallest γ range, "0.1-0.2", the selected methods that perform well during the split-sample test outperform the benchmark-to-beat. The better result for minimal γ ranges is probably partially related to the advantage of the tuning, which leads to more predictions of 0.1 within the regionalization. The benchmark-to-beat shows the best performance for γ values between 0.2 and 0.5. The good performance for basins with calibrated γ values between 0.2 and 0.5 is probably related to the benefit of using the logarithmical version of γ in the benchmark-to-beat, leading to more estimates of smaller values. However, this affects only 12 % of the basins, as calibrated values between 0.2 and 0.5 are not frequently present in the calibration result. Generally, the differences in KGE appear higher for smaller γ values, probably due to the decreasing parameter sensitivity with higher values (see Appendix B).

Given the variability in the performance of the regionalization methods across the depicted γ ranges, it is challenging to identify an overall best regionalization method using Fig. 4a. Therefore, we compare the various metrics of the KGE values depicted in Fig. 4b. The analyzed metrics are the minimum, maximum, mean, and median. Further, we count the number of poorly performing basins, defined as basins with a KGE below 0.2. In Fig. 4b, metrics that exceed the benchmark-to-beat are grey-shaded.



Method	Min	Median	Mean	Max	≤ 0.2
CAL (donor)	*0.402*	*0.679*	*0.672*	*0.939*	*0*
CAL (p.-ung.)	*0.403*	*0.674*	*0.663*	*0.953*	*0*
B2B	-1.060	0.627	0.587	0.944	17
MLR (best)	-0.708	0.633	0.606	0.951	22
MLR (worst)	-0.555	0.602	0.578	0.951	28
knn (best)	-0.955	0.626	0.597	0.953	18
knn (worst)	-2.937	0.604	0.545	0.926	37
SI (best)	-0.708	0.627	0.607	0.953	13
SI (worst)	-2.937	0.607	0.556	0.951	38
SP	-9.040	0.628	0.584	0.954	17

Figure 4: a) KGE values of pseudo-ungauged basins from split-sample test grouped by the range of calibrated γ values, b) selected metrics of KGE values from the pseudo-ungauged basins (better or equal performance to the benchmark-to-beat is highlighted in grey), and c) histogram of the number of pseudo-ungauged basins with a KGE below 0.2 and the corresponding number of methods exhibiting this performance loss.

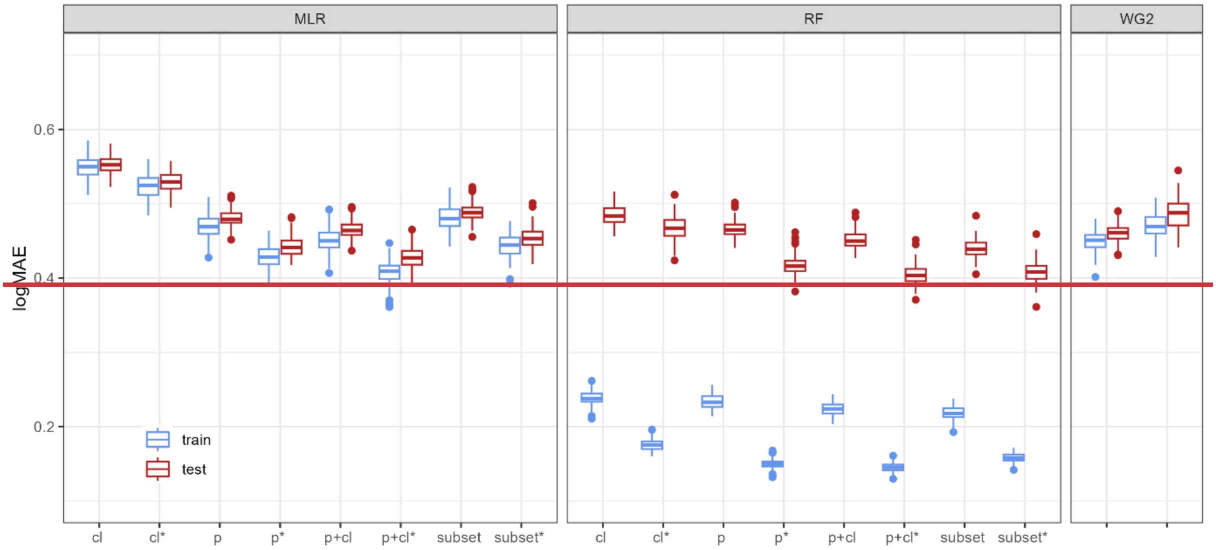
Comparing the KGE metrics in Fig. 4b reveals that the methods showing higher logMAE values in our split-sampling test ensemble also show lower performance in simulating discharge. For example, all mean (and median) KGE values of the "worst" methods are below the mean KGE of 0.587 from the benchmark-to-beat, ranging from 0.545 to 0.578. This indicates that the used logMAE between regionalized and calibrated values is a valid tool for a preliminary selection of adequate methods for the regionalization of WaterGAP3. However, for a more comprehensive analysis, we recommend additionally analyzing the accuracy of simulated discharges, as the logMAE of calibrated and regionalized parameter values simplifies the inherent complexity between model parameters and model performance.

Moreover, SI (best) outperforms the benchmark-to-beat in all listed metrics, reducing poorly performing basins and enhancing well-performing basins. MLR (best) performs very similarly to SI (best), yet it shows a higher number of basins with KGE values below 0.2. In comparison to the benchmark-to-beat, it outperforms four out of five criteria. The remaining well-performing methods, SP and knn (best), demonstrate superior or equal performance to the benchmark-to-beat in three out of five criteria. SP results in an equal number of poorly performing basins, and the minimal KGE value is lower than for the benchmark-to-beat. The knn (best) approach has a slightly worse median of KGE, i.e., -0.001, and one additional basin shows a KGE below 0.2.

As SI (best) outperforms the benchmark-to-beat in all metrics, we conduct a statistical test to ascertain whether there is a statistically significant difference in KGE results between the methods. To this end, we use a paired Wilcoxon rank sum test to test the null hypothesis of whether the KGE differs significantly in central tendency. A significance level of 0.05 and an adjusted p-value are applied to correct for multiple comparisons (using the correction after Benjamini & Hochberg (1995)). The results demonstrate that SI (best) outperforms all "worst" methods and the benchmark-to-beat. However, the null hypothesis for SP and the "best" options of knn and MLR cannot be rejected. Consequently, rather than identifying a single alternative to the benchmark-to-beat, we have identified four.

Notably, all regionalization methods lead to poorly performing basins, as evidenced by the range of basins with a KGE below 0.2, varying from 13 to 37. In Fig. 4c, we examine whether there are basins that all methods cannot regionalize, thereby indicating a general insufficiency of the regionalization methods for these basins. The histogram indicates that most poorly performing basins belong to a single regionalization method. The high number of basins, which cannot be estimated well by a single regionalization method, illustrates the diverse shortcomings of the methods. A single basin shows poor performance across all methods. This is a basin of the river El Platanito in Mexico. The calibrated γ value is about 1.5, and the corresponding KGE value in calibration is 0.466. This basin appears to be highly sensitive to γ, with an inaccuracy in the estimated γ having a significant impact on the accuracy of river discharge. For example, the benchmark-to-beat estimates γ to 1.0, which is close to the calibrated value of 1.5. However, the KGE value of the simulated discharge using the benchmark-to-beat is -0.158 due to a high overestimation of the variation and mean of the discharge. This high sensitivity seems outstanding and is likely attributable to the absence of waterbodies and snow, supporting a potentially high impact of γ on the model simulation (Kupzig et al., 2023) in conjunction with a relatively small basin size (ca. 6,600 km^2).

~~Here, we examine the traditional methods (MLR, SI, SP) by comparing the ensemble of MAEs from training and testing to each other and the benchmark-to-beat (see Fig. 3). Thus, applying knowledge of the optimal parameter space enhances the quality of regionalization. This positive effect is not surprising, as incorporating a-priori information about parameter distribution strengthens parameter estimation (e.g., described in Tang et al., 2016 using the Bayes Theorem).As for all traditional methods, there is no significant performance loss between training and testing, we will further focus on the performance in testing for evaluating the methods. When assessing the MLR and the SI approach, it becomes apparent that using only the climatic descriptors is insufficient for regionalization as it provides worse estimates than the benchmark-to-beat. The exclusive selection of physiographic descriptors (slope class, forest %, and permafrost %) performs better, and yields results comparable to our benchmark-to-beat for both methods. Using climatic and physiographic descriptors jointly increases the performance of SI by approximately 0.1 in median MAE. For MLR, the improvement is almost neglectable.~~



~~**Figure 3: Split sampling results for the benchmark-to-beat taken from WaterGAP2 (WG2) and different versions of the traditional regionalization methods: Multiple Linear Regression (MLR), Similarity Indices (SI) and Spatial Proximity (SP).**~~

~~Thus, using only climatic descriptors – in our case, the mean temperature and information about radiation – is insufficient for regionalization. Instead, physiographic descriptors appear more critical for regionalization than the selected climatic descriptors. However, the best results are obtained when combining climatic and physiographic descriptors. Others often apply the combination of climatic and physiographic descriptors, leading to optimal regionalization results (e.g., Oudin et al., 2008; Reichl et al., 2009).~~

~~The reduced importance of climatic descriptors is surprising, as the climatic descriptors tend to have a higher information gain and correlation to the model parameter (see Table 1). Moreover, climatic information is often used as a central part of other regionalization studies, e.g., to assess regionalization (e.g., Parajka et al., 2013; Guo et al., 2020). One possible reason for this discrepancy in other studies is that we used pure meteorological data as climatic descriptors for the regionalization method. In contrast, others used derived information such as Köppen-Geiger climate zones or the Aridity Index (e.g., Beck et al., 2016; Yoshida et al., 2022).~~

~~When expanding the analysis to all descriptors, the performance changes slightly (i.e., mean MAE +/- ~0.05). Thus, increasing the number of descriptors does not increase the performance of regionalization at some point (in line with Oudin et al., 2008 using a comparable Physical Similarity approach). This suggests that uncorrelated, non-redundant descriptors do not interfere with the regionalization using SI and MLR. Instead, a certain amount of information is beneficial to increase the regionalization method. After reaching this point, adding descriptors does not increase the performance, probably because all extractable information is already present in the given descriptors.~~

~~Using an ensemble of ten donor basins for the SI approach results in slightly better MAE values in most cases than applying a single donor basin (see Appendix B). More remarkably, the variation in the MAE values decreases significantly for all ensemble approaches (i.e., the reduction in standard deviation in MAEs is about 50%). Thus, introducing an ensemble approach for SI does not significantly improve the prediction performance. Still, it increases the likelihood that the prediction will perform well, i.e., be more robust. The positive effect of an ensemble approach for SI is already noted (Oudin et al., 2008). However, the literature-based number of donor basins might be adopted in future applications to be optimal for WaterGAP3, probably leading to higher performance.~~

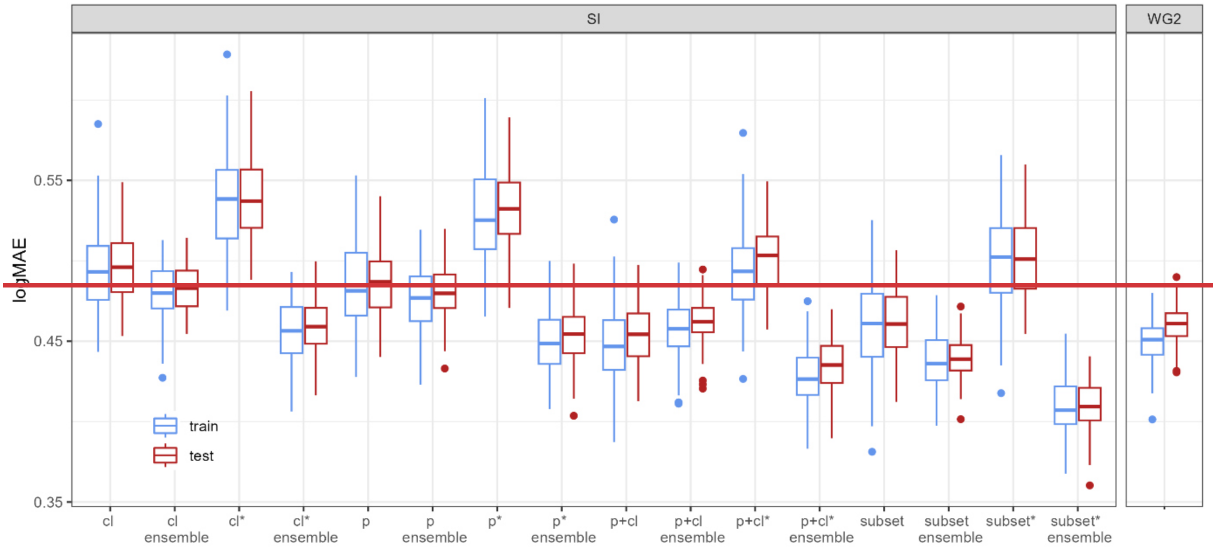
~~The introduction of tuning led to a significant increase in prediction performance for MLR, i.e., the median MAE for all MLR approaches improved by 0.04 (“cl”) and ~0.14 (others). For the ensemble-based SI approach, the tuning improves the median MAE by about 0.07 to 0.12. Thus, applying knowledge of the optimal parameter space enhances the quality of regionalization. This positive effect is not surprising, as incorporating a-priori information about parameter distribution strengthens parameter estimation (e.g., described in Tang et al., 2016 using the Bayes Theorem).~~

~~The SP approach is the simplest applied, evaluating distances to the centroids without requiring regression or clustering. Thus, there is no training performance, only a testing performance. Applying the approach leads to a median MAE of 1.356, which is better than the benchmark-to-beat (median MAE in the testing of 1.544) and has the same quality as the best MLR and SI approaches without tuning (median MAE of 1.394 and 1.367, respectively). The good performance of SP is in accordance with other studies (e.g., Oudin et al., 2008; Qi et al., 2020). It indicates that this simple approach is suitable for WaterGAP3.~~

~~Nevertheless, the well-performing SP on a global scale is surprising as the distances between basins are potentially large and hydrological processes may strongly vary. It is probably beneficial for the SP approach that γ comprises all kinds of errors, e.g., spatially localised errors in global forcing products (e.g., Beck et al., 2017 reported errors for arid regions in the precipitation product) or inaccurately represented processes for larger regions. Thus, the estimation of γ might be appropriate, but not because of the same hydrological behaviour but due to the same kind of errors.~~

~~3.2 Evaluating Machine Learning-based Approaches~~

~~In this section, we assess whether machine learning-based approaches outperform the benchmark-to-beat and are suitable as a new regionalization method for WaterGAP3. We compare the ensemble of MAE for training and testing for RF and k-means with the benchmark-to-beat (see Fig. 4).~~



~~Figure 4: Split-sampling results for the benchmark-to-beat taken from WaterGAP2 (WG2) and different versions of machine learning-based approaches: k-means (in combination with knn) and RandomForest (RF).~~

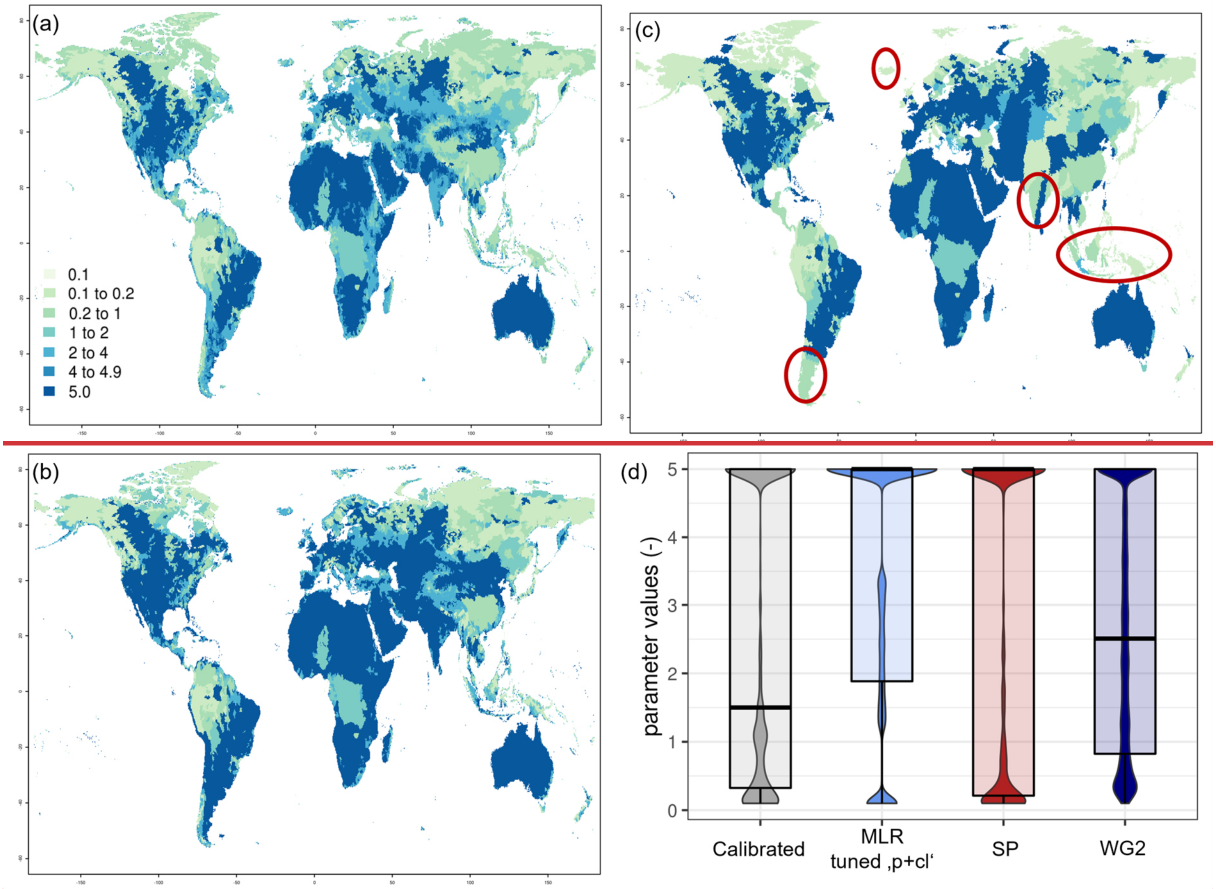
~~The RF approach is highly accurate within the training, i.e., fitting to calibrated γ values works well for gauged basins. However, it suffers a significant loss in performance when predicting the γ values for the pseudo-ungauged~~

~~basins. Although RF still has low MAE values in testing, the loss in performance from training to testing is significantly higher compared to other methods. This performance loss indicates that RF is not a robust regionalization method for WaterGAP3. Other studies which reported good performance of RF in terms of regionalization have not investigated the stability of the performance from training to testing (Golian et al., 2021; Wu et al., 2023). Likely, the mathematical problem of predicting the calibrated parameter for WaterGAP3, with all its challenges (e.g., tailored and heavy-tailed parameter space, incorporation of many sources of errors), cannot be adequately solved by RF. Thus, although RF is known to be especially robust among other machine learning-based techniques, it shows symptoms of over-parameterization, meaning that the algorithm is too flexible and adjusts to noise in the data, missing the underlying systematic. This lack of robustness is particularly disadvantageous since, for WaterGAP3, regionalization is applied globally, requiring regionalizing large parts of the world.~~

~~The k-means approach does not show such a performance loss between training and testing in almost all variants. The only variant with comparable performance loss is the "highly flexible" k-means approach. Interestingly, the "highly flexible" k-means approach was developed to emulate the same flexibility as in SI, which does not show such performance loss between training and testing. This difference in robustness indicates that the applied k-means algorithm does not extract the information from the descriptors as efficiently as the SI approach. The lack of efficient data use for some clustering methods in the context of regionalization has already been reported by Pagliero et al. (2019). This could also contribute to the presented the k-means falling behind the benchmark-to-beat. Therefore, we conclude that the developed clustering is inappropriate for regionalizing WaterGAP3.~~

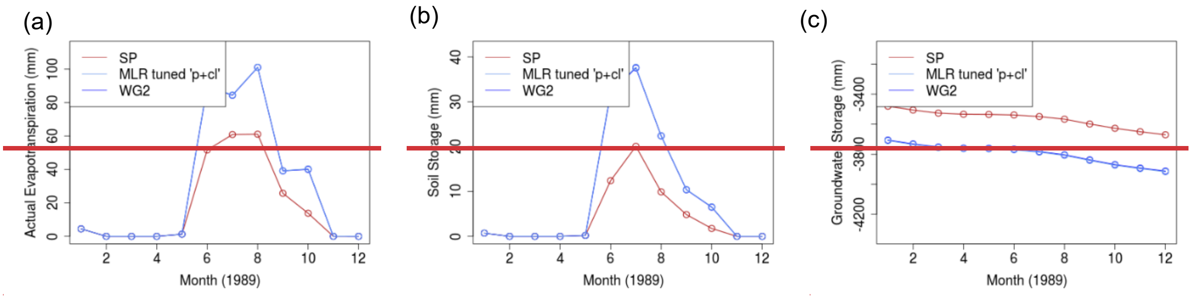
~~3.3 Implications of Regionalization~~

~~Finally, we highlight the possible implications of choosing regionalization methods for GHMs, where large parts of the world need to be regionalized. For this purpose, a local analysis of internal states and fluxes and a continental and global assessment of the water balance are undertaken. Therefore, we run WaterGAP3 from 1980 to 2016 with different γ distributions. We choose two equally valid solutions for the regionalization of WaterGAP3 to produce equally valid global γ distributions: (1) the SP approach because of its simplicity and because it outperforms our benchmark-to-beat, and (2) the tuned MLR "p+cl" because it outperforms our benchmark-to-beat and its application is very similar to the original regionalization approach of WaterGAP3. The tuned Similarity Indices "p+cl" with an ensemble of 10 donor basins is also a valid solution for regionalizing γ. However, its application is more complex than MLR and SP and differs considerably from the original WaterGAP3 regionalization. Therefore, it has not been implemented and tested. In addition, we run the model with our benchmark-to-beat as it is our reference for assessing changes. We use the best-performing benchmark-to-beat and MLR models out of the 100 trained models for the analysis.~~



~~**Figure 5: Global γ distribution for different regionalization methods, highlighting areas of differences (a) γ distribution using the MLR approach with parameter space tuning, using physiographic and climatic basin descriptors as independent variables, i.e., tuned MLR "p+cl", (b) benchmark-to-beat, WG2, (c) Spatial Proximity approach, i.e., SP and (d)** global distribution of regionalized and calibrated parameter values.~~

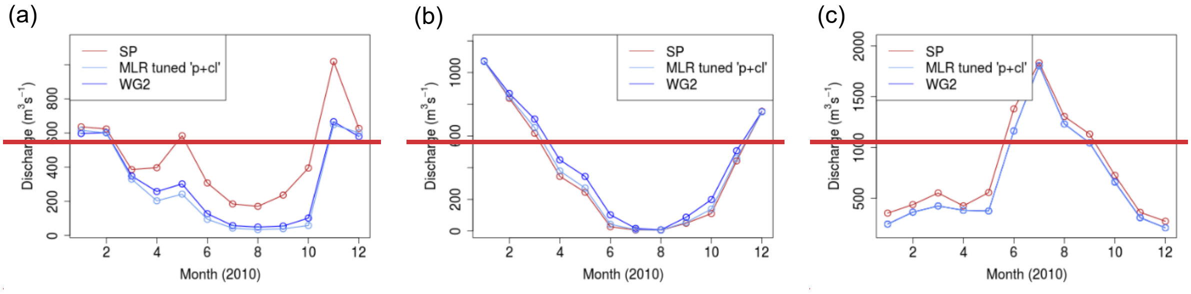
~~First, we compare the resulting global distribution of γ values for all three approaches (see Fig. 5). In particular, ungauged regions such as Indonesia, India and New Zealand exhibit significant differences in the predicted γ value. For these regions, the regionalized value varies depending on the methods used for regionalization. In contrast, ungauged areas such as North Africa do not differ much in regionalized values. Regionalization, therefore, appears to lead to a spatially varying uncertainty in ungauged regions. The differences in the regionalization methods also become apparent when comparing the resulting distribution of γ (see Fig. 5d). The approach MLR tuned "p+cl" tends to predict values at the upper bound more often than the other methods, which is probably due to the tuning within the method. The benchmark-to-beat approach from WaterGAP2 leads to a less heavy-tailed prediction than others. The SP-based approach shows the highest similarity to the distribution of the calibrated γ values.~~



~~**Figure 6: Differences in monthly internal states and fluxes of WaterGAP3 for one grid cell with varying regionalized value (SP: 0.325, MLR tuned "p+cl": 5 and benchmark-to-beat (WG2): 4.467243), located in India**~~

~~(21.519794°|70.566733°) for a) actual evapotranspiration, b) soil storage and c) groundwater storage for 1989 as an exemplary year. Note that MLR tuned “p+cl” and WG2 are so close that they appear to be one line.~~

~~To highlight the impact of local differences in the parameter value, we examine an exemplary location in India where the regionalized values are 0.325, 5 and 4.467243 for SP, MLR tuned “p+cl” and the benchmark-to-beat, respectively. We show the resulting actual evapotranspiration (AET), the filling of the soil storage and the groundwater storage for one exemplary year (see Fig. 6). The internal states and fluxes from the MLR tuned “p+cl” and the benchmark-to-beat are not significantly different for all states, as the two lines are very close and appear to be one single line. However, there are considerable differences between the two MLR-based approaches and SP, particularly in the amplitude of the AET and the soil storage. Acceleration effects cause the lower amplitudes for these two components. Reducing values of γ leads to a faster outflow of the soil storage, resulting in lower AET and soil moisture; additionally, smaller values of γ lead to higher groundwater storage due to accelerated percolation.~~



~~Figure 7: Simulated monthly runoff using three different regionalization methods for a) the Tiber, b) the Ebro and c) Rio Negro (in Argentina) for 2010 as an exemplary year.~~

~~Further on, we highlight the local effects of regionalization on discharge for the Tiber, the Ebro and Rio Negro for one exemplary year in Figure 7. Whereas the simulated discharge is higher for SP compared to the other methods in the Tiber and Rio Negro, the discharge is lower for the Ebro. Thus, one regionalization method does not always increase or decrease the discharge but results in locally varying effects on the water balance. Moreover, the similar results for MLR tuned “p+cl” and the benchmark-to-beat on the grid cell level (see Figure 6) propagate to a similar discharge pattern at the basin scale. Further, differences between SP and the other regionalization methods at the grid scale can lead to high differences at the basin scale, i.e., the simulated discharge of the Tiber is almost doubled for SP in May.~~

~~Finally, we evaluate how the observed variation due to different regionalization approaches propagates globally. Therefore, we assess the quantitative influence of regionalization by comparing a key component of the water balance, i.e., outflow to the ocean and inland sinks. Table 2 shows the resulting differences in the selected flow for all three model runs, aggregated to continental and global scales. The results highlight that the differences in mean annual outflow vary spatially and between the regionalization methods. The results of SP differ significantly from the two MLR-based approaches in some parts of the world. In Oceania, the SP approach exhibits a deviation of 7.7 % in the selected flow compared to the benchmark-to-beat. This difference may be attributed to the significant disparity in γ between the two methods in New Zealand (see Fig. 5).~~

~~Table 2: Mean outflow to the ocean and inland sinks in km^3 yr^{-1} between 1980-2010~~

~~Continent~~	~~benchmark-to-beat~~	~~MLR~~	~~SP~~
~~Africa~~	~~5005.10~~	~~0.972~~	~~0.968~~
~~Asia~~	~~15977.39~~	~~1.005~~	~~1.114~~
~~Oceania~~	~~1188.42~~	~~0.977~~	~~0.923~~
~~Europe~~	~~3028.47~~	~~0.981~~	~~1.030~~
~~South America~~	~~11612.39~~	~~0.997~~	~~1.039~~
~~North America~~	~~7283.21~~	~~0.994~~	~~1.025~~
~~Global~~	~~44094.97~~	~~43876.01~~	~~46456.35~~

~~Similarly, SP exhibits a high deviation of 11.4 % in the mean outflow in Asia, which is likely due to the variation of γ in India (see Fig. 5). In contrast, the southern part of South America, which shows a relatively high deviation in γ, does not lead to a significant deviation in the mean outflow for the continent. This limited impact of varying parameter values in southern South America may be attributed to the lower water availability in this region, which only slightly affects the continental water balance. These results suggest that the impact of regionalization methods on the continental water balance depends on (1) the variation in predicted parameter values and (2) the region's sensitivity to the water balance. Examining the global estimates, the differences between the benchmark-to-beat and SP results in approximately 2400 km^3 yr^{-1}, which is in the range of inter-model differences (see Table 2 in Widen-Nilsson et al.,2007).~~

~~Although the two newly developed methods performed similarly during the split-sample test, significant differences were observed when simulating the water balance. It was expected that the methods MLR tuned "p+cl" and SP methods would differ less due to their similar performance during the split-sample tests. However, it became apparent that the two MLR-based methods resulted in more closely simulation results than the SP-based approach. This indicates that the method selection, such as spatial proximity-based or regression-based, has a greater influence on the regionalization than the details of executing the method. Moreover, the split-sample test should be extended to get deeper insights into the method's robustness. For example, the SP and SI robustness check could be extended by the so-called "HDes" approach, which Lebecherel et al. (2016) recommended. In this approach, the closest basin to the corresponding (pseudo-) ungauged basin would be ignored during the regionalization to measure the robustness of the regionalization method.~~

3.4 Impacts on runoff simulations

To evaluate the impact of runoff simulations, we apply an ensemble of regionalization methods generating γ estimates for the worldwide ungauged regions. Within the ensemble, we use the four methods SI (best), knn (best), MLR (best), and SP that (1) outperform the benchmark-to-beat regarding the logMAE of regionalized and calibrated values and (2) perform similarly to each other and better than the benchmark-to-beat in KGE for monthly discharge. Additionally, we use the benchmark-to-beat as the fifth member of our regionalization method ensemble. The entire set of 933 gauged basins is used for regionalizing γ, resulting in five distinct worldwide distributions of γ. The spatially distributed standard deviation of the regionalized values is shown in Fig. 5.

In particular, the southern parts of South America, the northern and southern parts of North America, and Central Asia reveal differences in γ across the ensemble of regionalization methods (see Fig. 5). In Europe, the highest differences in regionalized values are observed in Italy, Great Britain, and northern Portugal. In Oceania, the highest values in standard deviation of γ are in Tasmania, New Zealand, and the southwest of Australia's coast. In contrast, a minor variation in γ is apparent in northern Africa, most parts of Australia, and the East of the Dead Sea. Thus, the uncertainty associated with globally regionalizing γ seem to vary across different regions.

~~4. Conc~~

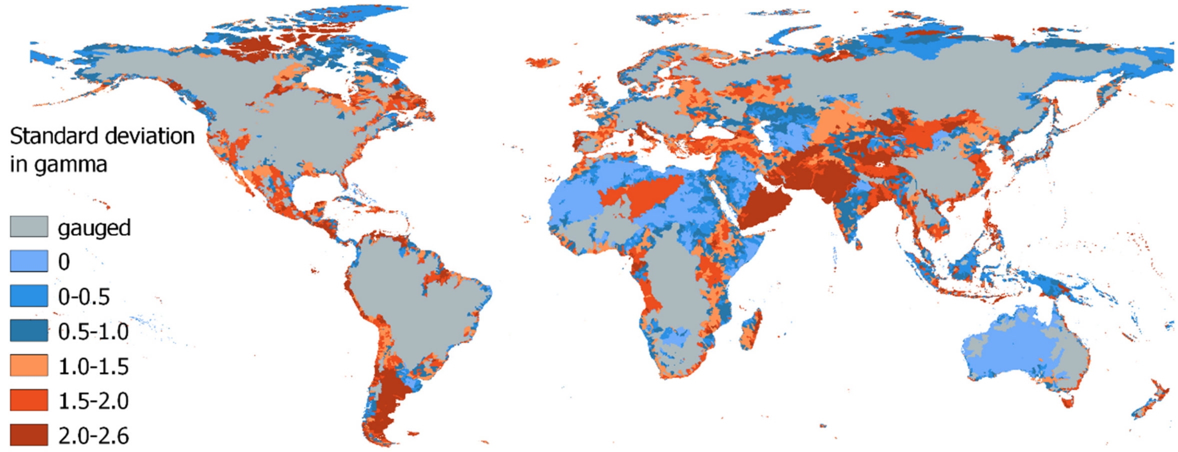


Figure 5: Standard deviation in regionalized γ values using the best approaches of MLR (best), SI (best), SP, knn (best), and the benchmark-to-beat. Note that dry regions without discharge are set to zero.

An example of how these uncertainties in regionalized values propagate through the water system is presented in Fig. 6. This figure displays the coefficient of variation of the mean yearly discharge between 1980 and 2016 based on the five simulation runs. Moreover, we highlight the effect on rivers in ungauged regions by showing the resulting seasonal pattern, i.e., the simulated long-term mean of monthly river discharge for three exemplary rivers. These rivers are the Río Bravo in Mexico, the Tiber in Italy, and the Tamar River in Tasmania. Each river is located in an ungauged region, where the standard deviation in γ is high (see Fig. 5).

Comparing Fig. 5 and Fig. 6 reveals that regions showing variability in γ tend to exhibit variation in mean yearly discharge. However, the impact of variation in γ on the simulated discharge appears to vary spatially. Some regions showing a high degree of variation in γ do not exhibit a correspondingly high degree of variation in discharge. For example, 45 % of all ungauged regions showing a low variation in discharge, i.e., the coefficient of variation is below 0.5, exhibit a standard deviation of more than one in γ. In contrast, about 89 % of the ungauged regions showing a higher discharge variation exhibit a standard deviation of more than one in γ. Thus, variation in γ does not necessarily lead to variation in river discharge, but it increases the likelihood that a region's discharge is affected. The spatially varying impact of γ is likely related to varying sensitivity regarding γ in the ungauged regions, which depends on numerous aspects, e.g., snow occurrence or waterbodies (see Kupzig et al., 2023).

About 11 % of the ungauged area exhibits variations in yearly river discharge exceeding 50 % of the mean. These regions are primarily in southern South America and Central Asia. A further 62 % of the ungauged area exhibits variations in yearly river discharge between 10 % and 50 % of the mean. These regions are mainly located on the northern coast of Russia and northern Canada, Indonesia, and Tasmania. Other areas, like most ungauged regions of Africa and Australia, show almost no impact, i.e., the variation in yearly discharge is less than 10 % of the mean. In northern Africa, one region exhibits higher values in the coefficients of variation. These values are attributable to minimal discharge values, resulting in comparatively high coefficients of variation in this region.

Considering the variation in the seasonality in the selected ungauged river systems (see Fig. 6b-d), the temporal impact of regionalization varies across the local landscape. For the Tamar River in Tasmania, as illustrated in Fig. 6d, the variation is higher at the start and end of the dry periods in October/November and April/May, respectively. The spread in monthly mean discharge is about 0.7 m^3s^{-1} to 1 m^3s^{-1} in these periods. The Tiber in Italy and the Río Bravo in Mexico exhibit a similar pattern: using the regionalized γ values of SP leads to much higher discharge rates than other ensemble members, introducing broad uncertainty bands. For the Tiber, this leads to seasonal estimates varying between 1.2 % (in January) and 11 % (in October) of the mean yearly sum. The Río Bravo shows variations in its seasonal pattern, with values ranging from 2.2 % (in February) to 6.8 % (in October) of the mean yearly sum. Thus, all rivers display a temporally varying impact. Whereas the main variation in the discharge of the Río Bravo and the Tiber is mainly attributed to the SP regionalization run, for the Tamaris River, all regionalization runs contribute to the varying long-term monthly mean in discharge.

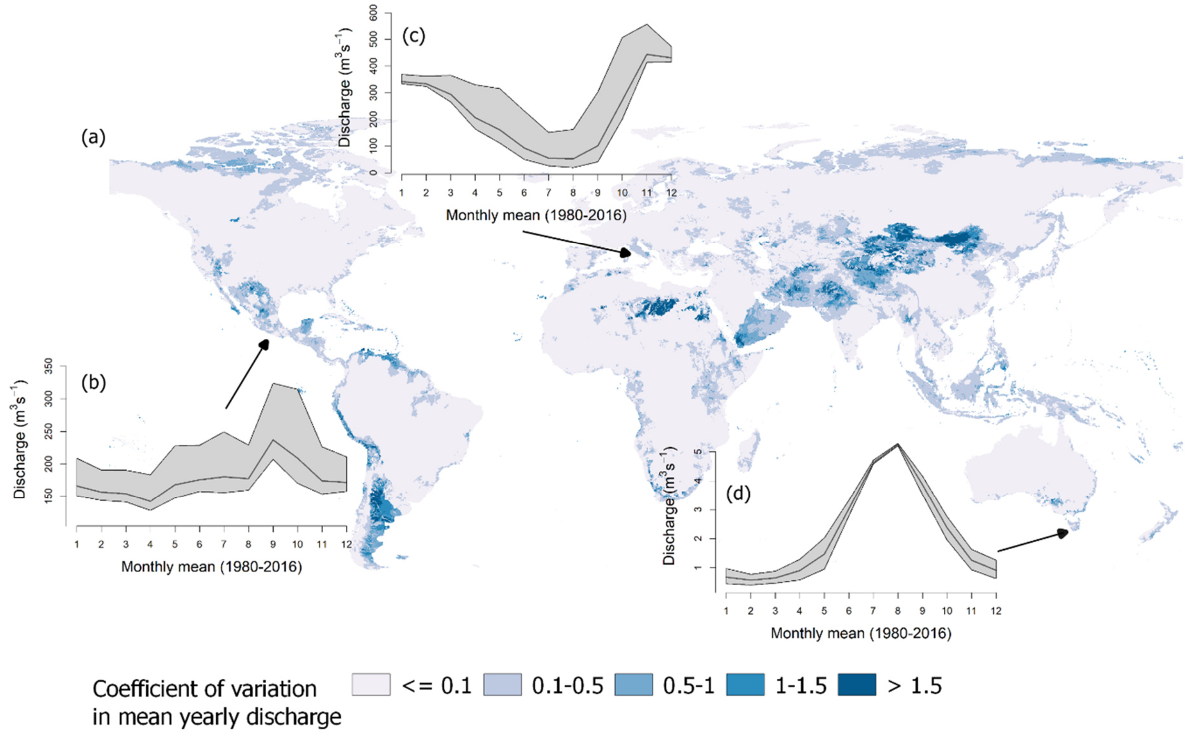
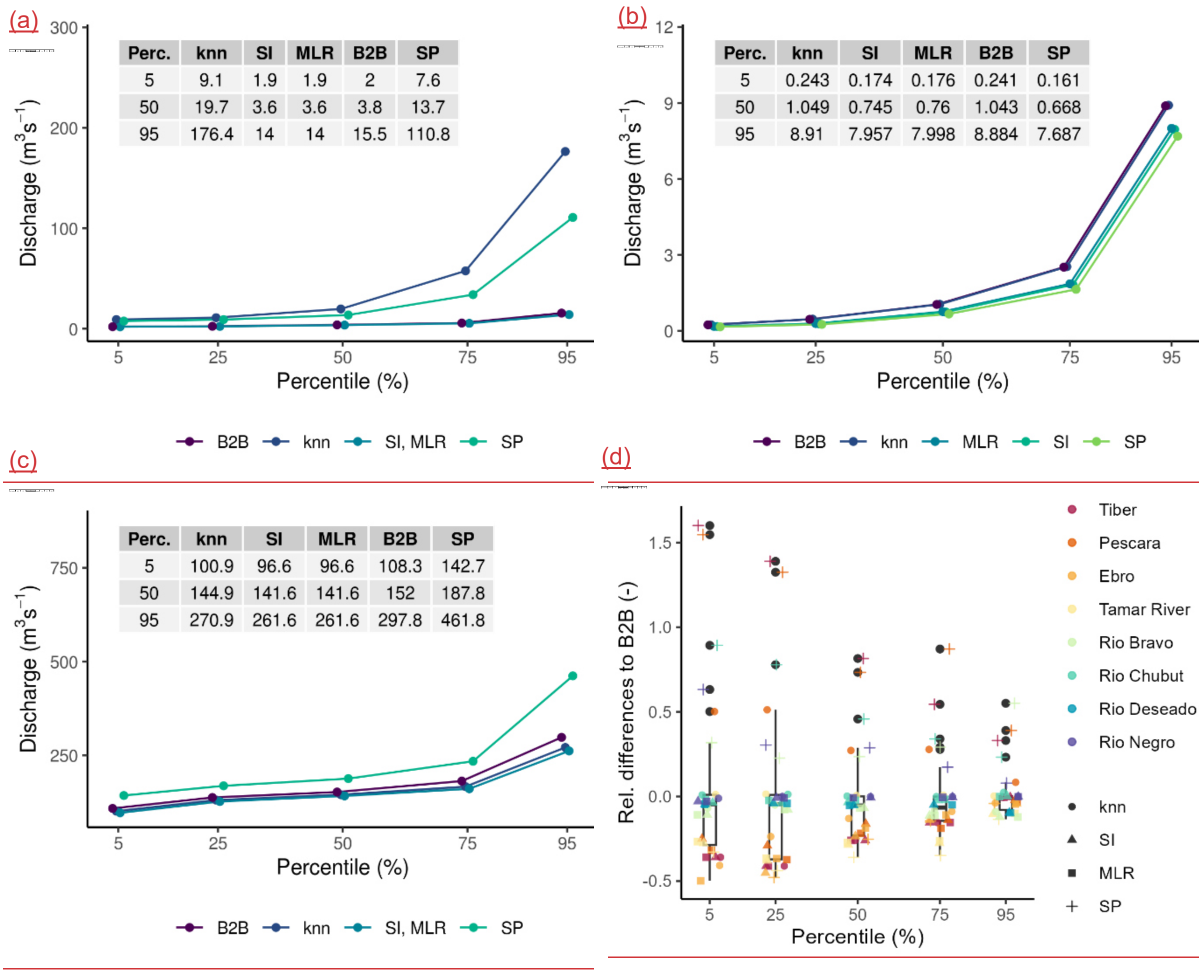


Figure 6: a) Global map of the coefficient of variation in mean yearly discharge for the applied regionalization methods. Resulting differences in the regionalization ensemble regarding the long-term mean of monthly discharge are depicted for: b) the Río Bravo in Mexico, c) the Tiber in Italy and d) the Tamar River in Tasmania. The grey-shaded area indicates the range of the long-term mean of monthly discharge and the black line indicates the mean off all simulation runs.

To gain a deeper understanding of the local impact of regionalization on runoff simulations, we analyze the annual percentiles from 1980 to 2016 for Río Deseado in Argentina, Río Bravo, and Tamar River, displaying the mean percentile of all years (see Fig. 7a-c). As the Tiber and Río Bravo display high similarities in the resulting patterns of percentiles, we demonstrate the impact by showing the percentiles from the Río Bravo. Additionally, we compare the relative differences in the mean for each percentile using eight ungauged river systems (see Fig. 7d), as previously done by Gudmundsson et al. (2012) for nine GHMs. To calculate the relative difference, we subtract the mean annual percentile of a method from the corresponding mean annual percentile of the reference and divide the resulting difference by the mean annual percentile of the reference. Instead of using observed flow as a reference, we use the annual percentiles of our benchmark-to-beat. As river discharge is already spatially aggregated information, it is unnecessary to spatially aggregate grid cells to create results comparable to those of Gudmundsson et al. (2012), who used cell runoff. The evaluated river systems are Río Chubut, Río Deseado, Río Negro, Río Bravo, Tamar River, Tiber, Pescara, and Ebro.



(a)

Perc.	knn	SI	MLR	B2B	SP
5	9.1	1.9	1.9	2	7.6
50	19.7	3.6	3.6	3.8	13.7
95	176.4	14	14	15.5	110.8

(b)

Perc.	knn	SI	MLR	B2B	SP
5	0.243	0.174	0.176	0.241	0.161
50	1.049	0.745	0.76	1.043	0.668
95	8.91	7.957	7.998	8.884	7.687

(c)

Perc.	knn	SI	MLR	B2B	SP
5	100.9	96.6	96.6	108.3	142.7
50	144.9	141.6	141.6	152	187.8
95	270.9	261.6	261.6	297.8	461.8

Figure 7: Mean annual percentiles between 1980 and 2016 of simulated discharge using an ensemble of regionalization methods. The river are a) Río Deseado, b) Tamar River, and c) Río Bravo. In d), the relative differences in mean annual percentiles to the benchmark-to-beat of eight ungauged river systems are presented. Negative values indicate smaller mean annual percentiles than the benchmark-to-beat. Note that all data points from Río Deseado for knn and SP are excluded as the values are above 2.0.

In Fig. 7a, Río Deseado is highly affected by uncertainties in simulated discharge due to the different regionalization methods; all segments of the percentiles show high variations where the absolute spread is increasing with increasing percentiles. For SP and knn (best), the discharge is highest, e.g., estimating a median discharge of 13.7 m^{3}s^{-1} and 19.7 m^{3}s^{-1}, respectively. For the other methods, the simulated discharge is low, e.g., SI and MLR result in an equal median discharge of 3.6 m^{3}s^{-1}. The Tamar River in Fig. 7b also shows increasing absolute differences between the methods for higher percentiles, with the benchmark-to-beat approach leading to the highest discharge. For the Río Bravo, the absolute differences between the highest result of SP and the other methods remain almost constant until the 75th percentile. For the 95th percentile, the absolute differences increase rapidly from about 40 m^{3}s^{-1} (75th percentile) to nearly 200 m^{3}s^{-1} (95th percentile). The exemplary results of Río Deseado and Río Bravo indicate a potentially high degree of uncertainty regarding the high percentiles in discharge simulation. These uncertainties put the results of global flood frequency analysis (e.g., Ward et al., 2013) in ungauged regions at risk as the time series of annual maxima might be even more uncertain. Thus, the results of flood frequency analysis should be carefully interpreted in ungauged regions as the impact of parameter regionalization may be significant.

Upon examination of the relative differences to the benchmark-to-beat for eight ungauged river systems, it becomes evident that the impact of regionalization methods varies between ungauged river systems (e.g., Río Negro exhibits almost no variation, but Ebro does). Moreover, it becomes apparent that some regionalization methods contribute more to the variation in estimated discharge than others. The methods contributing most are knn (best) and SP. For knn (best), 10 of the 40 relative differences are higher than |0.3|. For SP, even 29 out of the 40 relative differences are higher than |0.3|. The results of SI (best) and MLR (best) are very similar, indicating high similarity in performance. This is consistent with the KGE evaluation (see Chapter 3.3), in which they performed similarly. The observation in Fig. 7d that higher relative differences of discharge simulations occur in drier percentiles is also reported in Gudmundsson et al. (2012). Moreover, the relative differences between the five regionalization runs seem comparable to the inter-model differences depicted in Gudmundsson et al. (2012), indicating the high impact of regionalization methods on the evaluated ungauged river systems.

Finally, Table 3 presents the estimated yearly mean runoff to the ocean for all five ensemble members. All estimates of global "runoff to ocean" range from 45,622 (SI (best)) to 47,069 (SP). Thus, the differences are on the scale of smaller inter-model differences (see Table 2 in Widen-Nilsson et al.,2007). The impact of regionalization becomes even more evident using an unsuitable regionalization method for WaterGAP3. For instance, the tuned kmeans ("subset") approach results in 42,862 km^3 yr^{-1} "runoff to ocean", increasing the spread between the methods to 4,208 km^3 yr^{-1} being in the scale of inter-model differences. This high impact of regionalization on global "runoff to ocean" is surprising, given that only 27 % of the world is ungauged, using the GRDC database. From this 27 %, most regions are in Australia and Africa, where minimal runoff is produced. In studies employing disparate models, e.g., for inter-model comparison, all regions are simulated in disparate ways.

The most significant deviations in the continental sums of "runoff to ocean" in Table 3 are due to SP. Only for Europe is the highest deviation related to MLR (best), not SP. Interestingly, the estimated sums of SP occasionally define the lowest and occasionally the highest extremes for the continents, lacking a systematic pattern. The outstanding role of SP is consistent with previous evaluations in this Chapter, where SP frequently contributes most to the variation in discharge. This suggests that SP may not be suitable for the global scale. Nevertheless, the pseudo-ungauged basins in the split-sample tests may also exhibit considerable distances from the observed basins. Given that SP achieved satisfactory results in both evaluations, using either the logMAE or the KGE, the evaluation indicates the method's suitability on a global scale. Thus, in the future, the split-sample test must be extended to gain deeper insights into the method's robustness and make a definitive statement about the method's suitability on a global scale. For example, the so-called "HDes" approach, recommended by Lebecherel et al. (2016), could be applied for this purpose. In this approach, the closest basin to the corresponding (pseudo-) ungauged basin is excluded from the regionalization process, thereby enabling an assessment of the method's robustness.

Table 3: Mean outflow to the ocean and endorheic basins in km³ yr^{-1} between 1980-2016. The highest continental deviation to the benchmark-to-beat is indicated in bold.

Runoff to ocean[1]	B2B	SI (best)	knn (best)	MLR (best)	SP
Oceania	1,127	-1.80 %	-2.20 %	-3.40 %	**-6.60 %**
Europe	3,098	-2.30 %	-0.10 %	**-2.60 %**	0.20%
Asia	16,676	3.50 %	0.30 %	1.60 %	**5.50 %**

Africa	5,203	-1.00 %	0.70 %	-0.30 %	**-3.60 %**
North America	7,517	0.30 %	1.00 %	-1.70 %	**2.20 %**
South America	12,032	1.30 %	1.40 %	-0.20 %	**4.90 %**
global	45,653	46,273	45,953	45,622	47,069

[1]including endorheic basin

Conclusion

Valid simulation results from GHMs, such as WaterGAP3, are crucial for detecting hotspots or studying patterns in climate change impacts. However, the lack of worldwide monitoring data makes adapting GHMs' parameters for valid global simulations challenging. Therefore, regionalization is necessary to estimate parameters in ungauged basins. This study applies regionalization methods for the first time to WaterGAP3, aiming to provide insights into selecting suitable regionalization methods and evaluating their impact on the runoff simulations. Traditional and machine learning-based methods are tested to assess the application of several regionalization techniques on a global scale. The concept of benchmark-to-beat and an ensemble of split-sampling tests are employed for a comprehensive evaluation. Moreover, the impact on runoff simulation is assessed using a wide range of temporal and spatial scales, i.e., from the daily to the yearly and from the local to the global scale.~~Valid simulation results from GHMs, such as WaterGAP3, are crucial for detecting hotspots or studying patterns in climate change impacts. However, the lack of worldwide monitoring data makes adapting GHMs' parameters for valid global simulations challenging. Therefore, regionalization is necessary to estimate parameters in ungauged basins. This study introduces novel regionalization methods for WaterGAP3 and aims to provide insights into selecting a suitable regionalization method and evaluating its impact on the simulation results. Traditional and machine learning-based methods are tested to assess the advantages of using new techniques on a global scale. The concept of benchmark-to-beat and an ensemble of split-sampling tests are employed for a comprehensive evaluation.~~

In this study, four regionalization methods outperform the benchmark-to-beat and thus are considered appropriate for WaterGAP3. These methods span the complete range of methodologies, i.e., regression-based methods and methods using the concept of physical similarity and spatial proximity. Moreover, the methods vary in the descriptors used to achieve optimal results. This highlights that different methods use descriptor sets with varying efficiency. All methods perform best when using climatic and physiographic descriptors, indicating that combining climatic and physiographic descriptors is optimal for regionalizing worldwide basins. Although random forest is known to be especially robust among other machine learning-based techniques, it shows symptoms of over-parameterization, indicating that the algorithm is too flexible and adjusts to noise in the data, missing the underlying systematic pattern.

Our results demonstrate that variation in the regionalized parameter value does not necessarily lead to variation in river discharge. However, it increases the likelihood that a region's runoff is affected. This spatially varying impact of γ is likely related to the varying sensitivity in ungauged regions regarding γ. Southern South America is a region identified to be especially sensitive to variation in γ. Furthermore, local effects on runoff simulations indicate a temporally varying impact. For example, some impacted rivers indicate a high degree of uncertainty regarding the high percentiles in discharge simulation. These uncertainties potentially lead to a significant impact on flood frequency analysis on a global scale, where the lack of gauging stations in certain regions calls for regionalization. The global impact of regionalization methods that perform well for WaterGAP3 appears to be in the order of minor inter-model differences. This impact rigorously increases when using a poorly performing method for WaterGAP3, underscoring the importance of carefully selecting regionalization methods.

The spatial proximity approach contributes most to the variation in estimated runoff. The outstanding role of this approach suggests that it may not be suitable for the global scale. However, as the pseudo-ungauged basins in the split-sample tests may also have considerable large distances to the observed basins, and the method achieves satisfactory results in all executed evaluations, it is not possible to make a definite statement about the method's suitability for the global scale. Further research is required to gain deeper insights into the methods' robustness, e.g., by extending the analysis by applying the recommended "HDes" approach (Lebecherel et al., 2016).

~~Our results suggest that the basin descriptor selection may not be crucial for regionalization in WaterGAP3 as long as a subset of the selected descriptors contains relevant information. Additionally, introducing an ensemble approach for Similarity Indices does not necessarily improve the prediction performance but increases the likelihood of robust predictions. Interestingly, the simplest regionalization method (using the concept of spatial proximity) outperforms most of the developed regionalization methods and the benchmark-to-beat. In contrast, the more complex, machine learning-based approaches deliver insufficient prediction performance. The inadequate performance may be attributed to an inefficient extraction of available information content from the descriptors and the blurring relationship between the calibration parameter and basin descriptors, which is caused by including multiple error sources in the calibration parameter values. This blurring relationship probably poses a high risk of over-parameterization, which hinders the use of more flexible machine learning-based approaches.~~

~~Regionalization appears to result in spatially varying uncertainty for ungauged regions, with India and Indonesia being particularly affected by higher uncertainty. The local impacts of regionalization in ungauged areas propagate to the global scale, where the water balance component “outflow to the ocean and inland sinks” changed by about 2400 km^3 yr^{-1}, which is in the scale of inter-model differences. As the selected regionalization method influences the regionalization more than details on the execution of the method, we recommend employing simulation runs that use multiple regionalization methods to account for the uncertainty induced by the chosen regionalization method. Considering the uncertainty induced by regionalization is especially important when analysing regions with a significant proportion of ungauged basins or high sensitivity to the examined target.~~

Code and data availability. The data and the supporting R-Code to reproduce this study's findings are available at https://doi.org/10.5281/zenodo.11833447~~DOI 10.5281/zenodo.10803089~~.

Authors contribution. JK developed, designed, and drafted the study. NK helped to design the experiment. MF provided feedback throughout the entire process and supported the writing.

Competing interests. The authors declare that they have no conflict of interest.

Appendix A: Global Map of derived global soil moisture storage

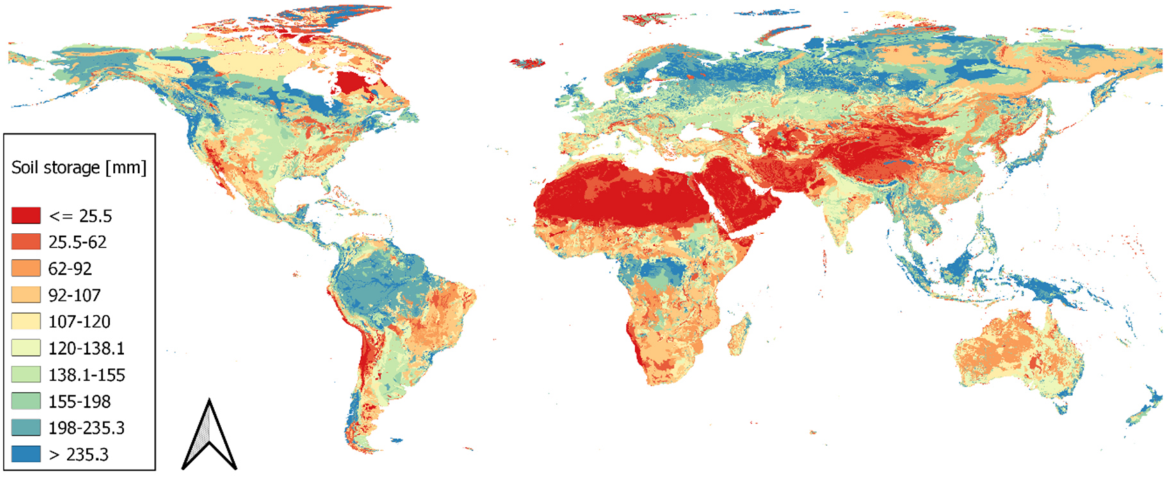


Figure A1: Global map of the size of soil storage based on Batjes (2012) and land use information (derived from Friedl & Sulla-Menashe, 2019)

Appendix B: Further analysis regarding the clustering of parameter values at the extremes

The clustered calibrated parameter values at the extremes of the valid parameter space (see Fig. 1b) are a known problem within the calibration. As the parameter space, i.e., the parameter bounds, is crucial for calibration and, in consequence, for regionalization, we address this issue by a brief sensitivity analysis to demonstrate that the clustering of the calibrated parameter values is more an issue of missing processes (or using additional parameter values) than an issue of inappropriate parameter space. As the lower limit of the calibrated parameter (0.1) is sufficiently small in comparison to other studies using a similar HBV-based approach for runoff generation processes (e.g., see the beta in Table A2 in Jansen et al., 2022), we focus on the sensitivity analysis on the upper limit of γ (5.0).

In the sensitivity analysis regarding the upper limit of γ, we applied the model formula (see equation B1) containing the model's parameter γ and modified it within the bounds of 0.1 and 10. Additionally, we modified the soil saturation varying from 1 % to 95 %.

$$outflow = precipitation_{effective} \cdot soil\ saturation^{gamma} \qquad \text{(B1)}$$

The calculated outflow and its relationship to the soil saturation and γ are depicted in Fig. B1 and B2. The incoming effective precipitation is defined as constant. As it is a factor in equation B1,, the results regarding incoming effective precipitation are linearly scalable.

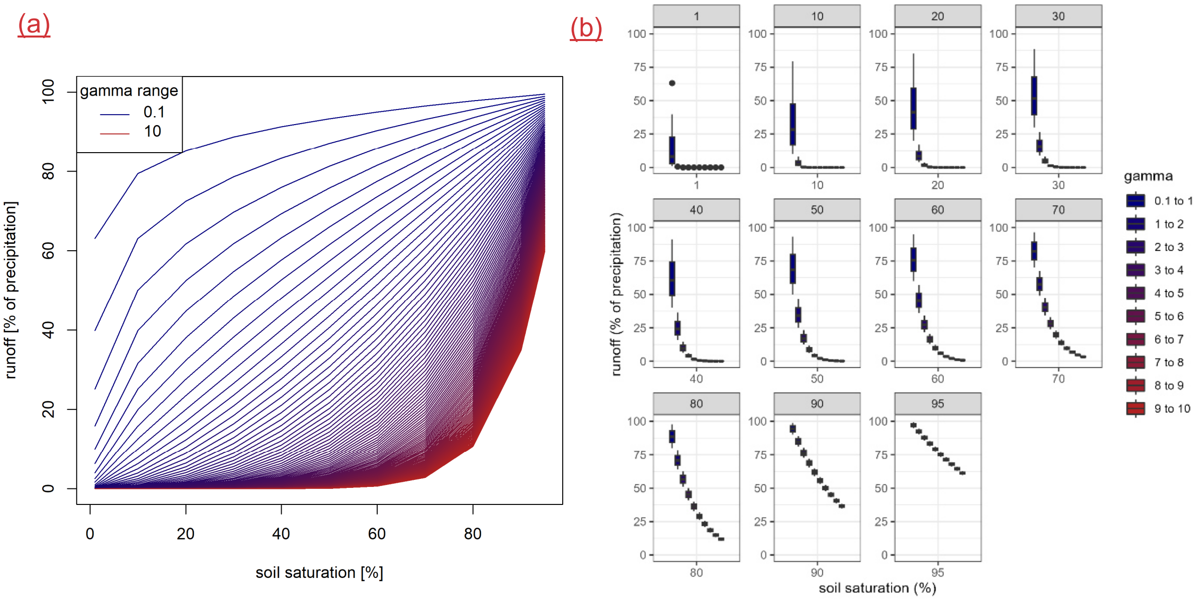


Figure B1: a) Runoff generation in the soil layer (neglecting overflow and evapotranspiration) using different values for the calibration parameter and increasing the soil-moisture, b) runoff generation for varying soil moisture grouped in bins of size one.

In the depicted Fig. B1, the runoff generation process differences between differing γ values become more linear when soil saturation increases. Thus, the non-linear model parameter becomes less critical for high soil moisture. Generally, the runoff generation process differences for higher γ values are more pronounced for higher soil moisture. For lower soil moisture, the smaller values have higher effects on the generated runoff. For example, for 70 % soil moisture, the differences for γ values ranging from 5 to 10 are between 3 % and 16 %. For the same soil moisture, the range in runoff generation varies from 16 % to 70 % for γ values between 1 and 5.

High γ values usually occur in dry regions (see Fig. 4b in Müller Schmied et al., 2021). In dry regions, high soil moisture values are not expected to occur frequently (e.g., see Khosa et al., 2020; Oloruntoba et al., 2024 for estimated and measured soil moisture in Africa and Draper et al., 2008 for estimated and measured soil moisture in Australia). It is, therefore, unlikely that higher γ values will significantly enhance the calibration result or decrease the issue of clustered calibrated parameter values at the higher end of the parameter space. More likely, the clustering of calibrated parameter values will be resolved in dry regions by incorporating additional (missing) model processes, such as evaporation from rivers or inaccurate representation of groundwater processes (Eisner, 2016, p. 49). Thus, the parameter bounds of γ (e.g., also used in Eisner 2016, p. 16; Müller Schmied et al., 2021; Müller Schmied et al., 2023) are not changed in this study.

Appendix C~~A~~: Basin descriptors

Overview of basins descriptors used in this study. All basin descriptors are derived from the original model input and aggregated with a simple mean method to basin values to produce the same spatial resolution as the calibrated model parameter.

- *Soil Storage*: The size of the soil storage, i.e., the maximal water content in the soil reachable for plants in ~~millimetres~~mm. The information is the product of rooting depth (defined in a look-up table) and the total available water content derived from Batjes (2012~~3~~).
- *Open Water Bodies*: The fraction of the area covered with open water bodies in the basin is given as a percentage. The model input is based on the GLWD database (Lehner & Döll, 2004).
- *Wetlands*: The fraction of area covered with wetlands in a basin is given in percentage. The model input is based on the GLWD database (Lehner & Döll, 2004).
- *Size*: Size of a basin in km^2.
- *Slope*: The mean slope class is calculated as described in Döll & Fiedler (2008) and based on GTOPO30 (USGS EROS data centre).
- *Altitude*: The mean altitude of a basin is given in ~~metres~~ meters above sea level and based on GTOPO30 (USGS EROS data centre).
- *Forest*: The mean fraction of the area covered with forest is given in percentage and derived from MODIS data (Friedl & Sulla-Menashe, 2019), where 2001 is used as a reference. All grid cells having a dominant International Geosphere-Biosphere Programme (IGBP) classification between one and five are defined as ~~“~~"forest~~”.~~".
- *Sealed Area*: The mean fraction of sealed area is given in percentage and derived from MODIS data (Friedl & Sulla-Menashe, 2019), where 2001 is used as a reference. All grid cells having an IGBP classification equal to 13 are defined as they would contain 60% of the sealed area. Note: The different treatment of forest and sealed area is based on the required model input; whereas the land cover is a classified value, the sealed area is a floating-point value.
- *Permafrost & Glacier*: The mean coverage of permafrost and glacier in a basin is given in percentage. It is based on the World Glacier Inventory and the Circum-Arctic Map of Permafrost and Ground-Ice Conditions.
- *Mean Temperature*: The mean air temperature is based on the meteorological forcing used to drive the model (Lange, 2019) covering the period 1979 to 2016 and given in degrees Celsius.
- *Yearly Precipitation*: The yearly precipitation sum is based on the meteorological forcing used to drive the model (Lange, 2019) covering the period 1979 to 2016 and given in ~~millimetres~~mm.
- *Yearly Shortwave Downward Radiation*: The yearly shortwave downward radiation is based on the meteorological forcing used to drive the model (Lange, 2019) covering the period 1979 to 2016 and given in Wm^{-2}.

The correlation between the defined basin descriptors is shown in Fig. A1. The variation within each basin descriptor for basins used for ~~regionalization~~ regionalization is shown in Fig. A2.

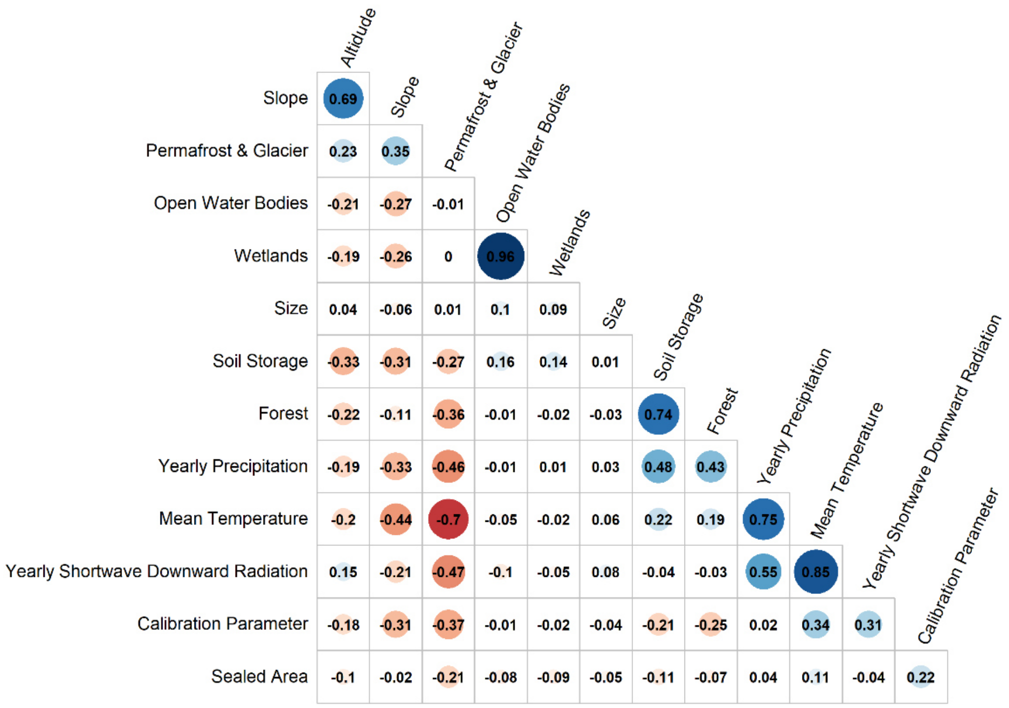


Figure C~~A~~1: Correlation between basins descriptors.

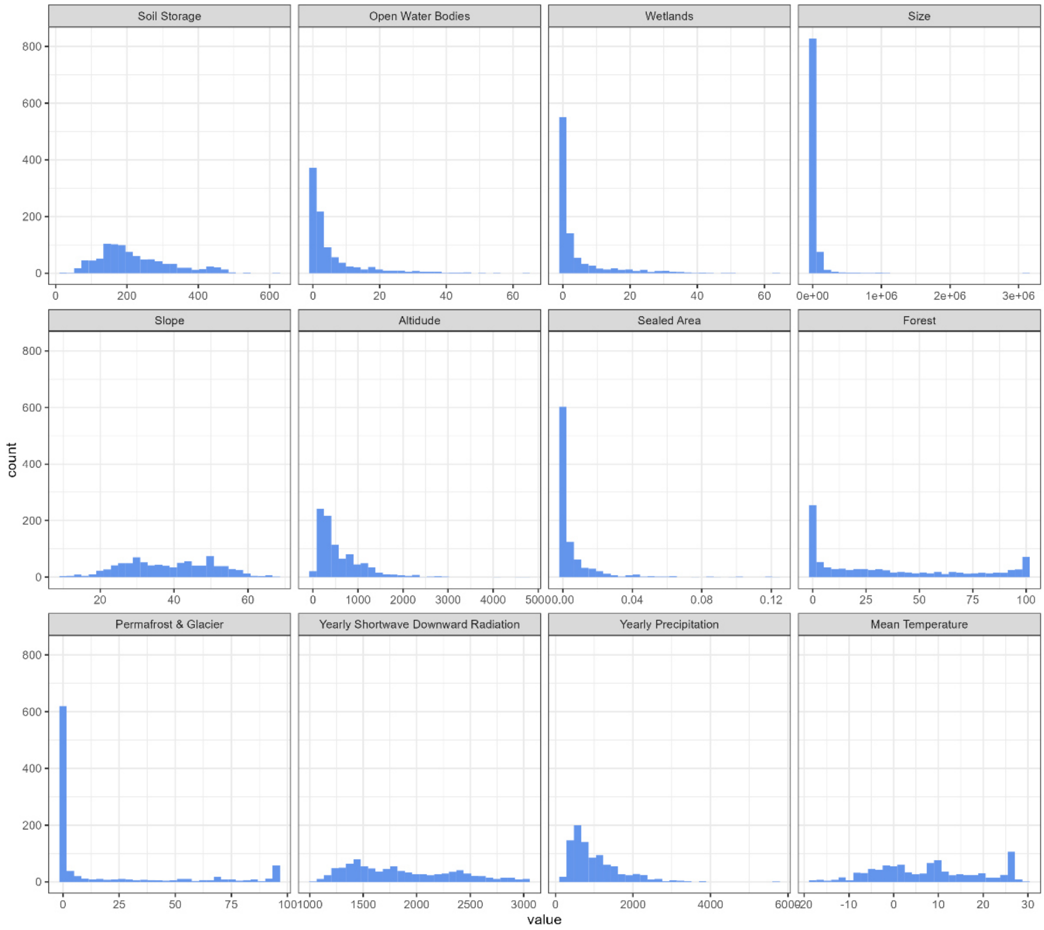


Figure C~~A~~2: Distribution of basins descriptors within all basins used for ~~regionalization~~ regionalization (n=~~1,236~~933)

~~Appendix B: Results of split-sample tests~~

~~Table B1: Summarized results of the split-sample tests for all regionalization methods~~

~~input~~	~~method~~	~~train (median)~~	~~train (sd)~~	~~test (median)~~	~~test (sd)~~
~~-~~	~~WG2~~	~~1.527~~	~~0.042~~	~~1.544~~	~~0.046~~
~~-~~	~~SP~~	~~-~~	~~-~~	~~1.356~~	~~0.057~~
~~el~~	~~MLR~~	~~1.474~~	~~0.039~~	~~1.485~~	~~0.019~~
~~p~~		~~1.871~~	~~0.034~~	~~1.881~~	~~0.015~~
~~p+el~~		~~1.457~~	~~0.038~~	~~1.473~~	~~0.018~~
~~all~~		~~1.394~~	~~0.039~~	~~1.425~~	~~0.024~~
~~el~~	~~MLR_t~~	~~1.322~~	~~0.040~~	~~1.331~~	~~0.027~~
~~p~~		~~1.830~~	~~0.041~~	~~1.843~~	~~0.030~~
~~p+el~~		~~1.307~~	~~0.042~~	~~1.337~~	~~0.030~~
~~all~~		~~1.245~~	~~0.042~~	~~1.292~~	~~0.034~~
~~el~~	~~RF~~	~~0.688~~	~~0.026~~	~~1.401~~	~~0.029~~
~~p~~		~~0.741~~	~~0.027~~	~~1.579~~	~~0.032~~
~~p+el~~		~~0.620~~	~~0.020~~	~~1.312~~	~~0.025~~
~~all~~		~~0.624~~	~~0.021~~	~~1.346~~	~~0.023~~
~~el~~	~~RF_t~~	~~0.465~~	~~0.020~~	~~1.310~~	~~0.039~~
~~p~~		~~0.494~~	~~0.023~~	~~1.540~~	~~0.042~~
~~p+el~~		~~0.378~~	~~0.017~~	~~1.183~~	~~0.037~~
~~all~~		~~0.345~~	~~0.014~~	~~1.181~~	~~0.034~~
~~el~~	~~SI_1~~	~~1.477~~	~~0.080~~	~~1.492~~	~~0.056~~
~~p~~		~~1.651~~	~~0.086~~	~~1.661~~	~~0.063~~
~~p+el~~		~~1.380~~	~~0.066~~	~~1.375~~	~~0.050~~
~~all~~		~~1.367~~	~~0.069~~	~~1.390~~	~~0.064~~
~~el~~	~~SI_10~~	~~1.398~~	~~0.046~~	~~1.397~~	~~0.029~~
~~p~~		~~1.558~~	~~0.047~~	~~1.556~~	~~0.027~~
~~p+el~~		~~1.326~~	~~0.044~~	~~1.321~~	~~0.025~~
~~all~~		~~1.398~~	~~0.049~~	~~1.402~~	~~0.028~~
~~el~~	~~SI_10_t~~	~~1.281~~	~~0.053~~	~~1.281~~	~~0.043~~
~~p~~		~~1.497~~	~~0.050~~	~~1.487~~	~~0.037~~
~~p+el~~		~~1.206~~	~~0.048~~	~~1.201~~	~~0.040~~
~~all~~		~~1.286~~	~~0.053~~	~~1.296~~	~~0.039~~
~~el~~	~~k-means~~	~~1.689~~	~~0.038~~	~~1.699~~	~~0.018~~
~~p~~		~~1.910~~	~~0.051~~	~~1.918~~	~~0.039~~
~~p+el~~		~~1.632~~	~~0.046~~	~~1.648~~	~~0.022~~
~~all~~		~~1.642~~	~~0.044~~	~~1.638~~	~~0.025~~
~~el~~	~~k-means_t~~	~~1.474~~	~~0.111~~	~~1.519~~	~~0.088~~
~~p~~		~~1.909~~	~~0.055~~	~~1.918~~	~~0.040~~
~~p+el~~		~~1.399~~	~~0.070~~	~~1.425~~	~~0.053~~
~~all~~		~~1.426~~	~~0.068~~	~~1.417~~	~~0.051~~
~~el~~	~~k-means flexible~~	~~1.065~~	~~0.048~~	~~1.553~~	~~0.097~~
~~p~~		~~1.191~~	~~0.046~~	~~1.991~~	~~0.142~~
~~p+el~~		~~0.982~~	~~0.040~~	~~1.568~~	~~0.125~~
~~all~~		~~0.957~~	~~0.044~~	~~1.515~~	~~0.114~~

Appendix D: Results of the ensemble of the split-sample tests

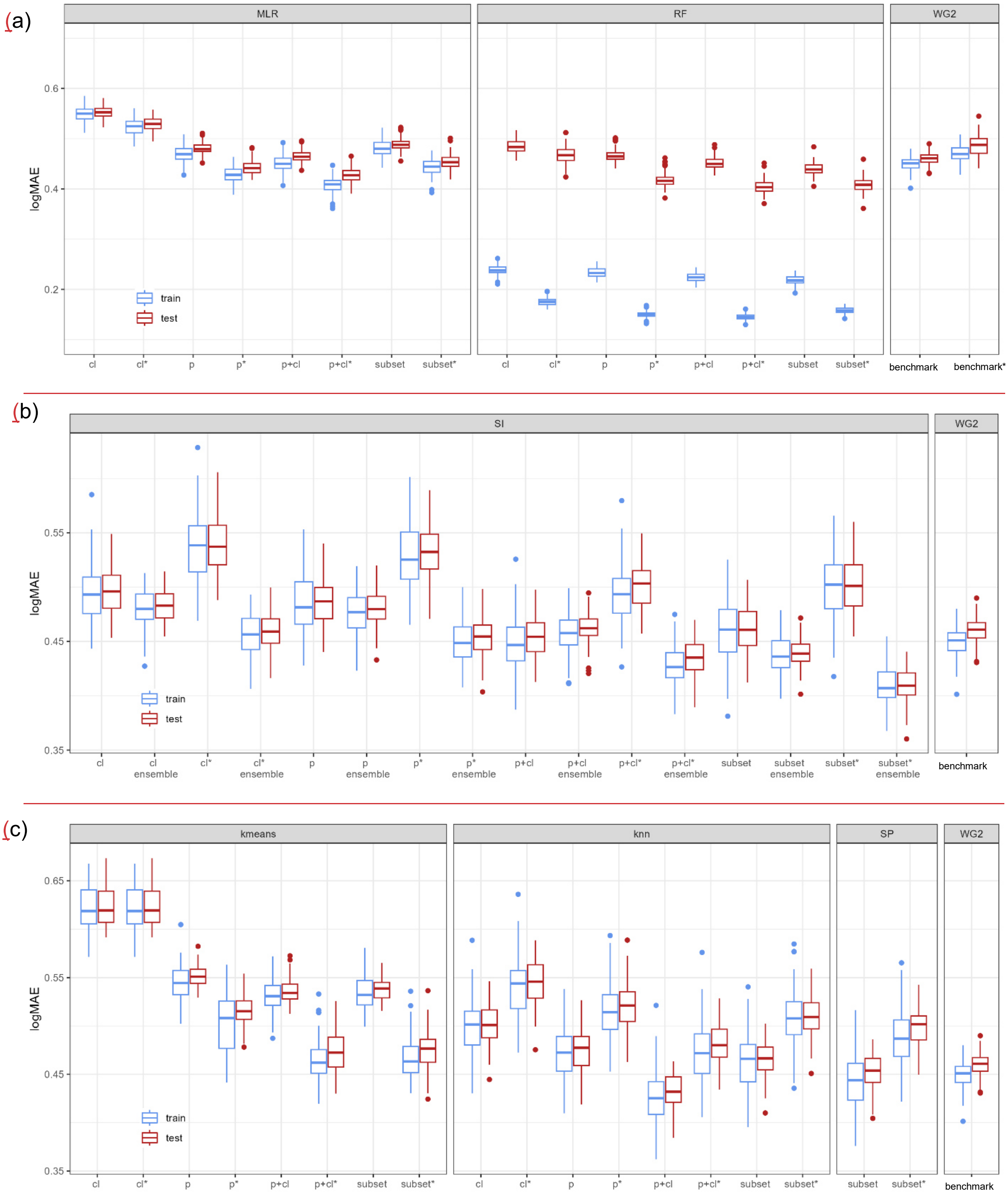


Figure D1: logMAE values for all 100 split-sampling tests using all variants of a) MLR, RF, and benchmark-to-beat, b) SI, and c) kmeans, knn, and SP. Note that the asterisk * indicates the tuned version of the method.

Table D1: Performance loss in median logMAE of the ensemble of split-sample tests from training to testing expressed in % of logMAE in training.

test (% train)	MLR	RF	SI		kmeans	knn	SP	B2B
			no ens.	ensemble				
cl	100.4	202.9	100.6	100.6	100	100	102.3	102.2
p	102.1	199.6	101.2	100.6	101.3	101.1		
p+cl	103.1	207.1	101.6	100.9	100.6	95.6		
subset	101.7	223.9	100	100.7	101.3	100.2		

test* (% train*)	MLR	RF	SI		kmeans	knn	SP	B2B
			no ens.	ensemble				
cl	100.8	266.9	99.8	100.7	100	100.4	103.1	104.1
p	103	277.3	101.3	101.3	101.4	101.4		
p+cl	104.4	277.9	102	102.1	102.2	101.7		
subset	102	258.2	99.8	100.5	103	100.2		

References

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., & Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrology and Earth System Sciences, 24(2), 535–559. https://doi.org/10.5194/hess-24-535-2020, 2020.

Arsenault, R, & Brissette, F. P.: Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches, Water Resources Research, 50, 6135–6153, https://doi.org/10.1002/2013WR014898, 2014.

Ayzel, G. V., Gusev, E. M., & Nasonova, O. N.: River runoff evaluation for ungauged watersheds by SWAP model. 2. Application of methods of physiographic similarity and spatial geostatistics, Water Resources, 44(4), 547–558, https://doi.org/10.1134/S0097807817040029, 2017.

Barbarossa, V., Bosmans, J., Wanders, N., King, H., Bierkens, M. F. P., Huijbregts, M. A. J., & Schipper, A. M.: Threats of global warming to the world's freshwater fishes, Nature Communications, 12(1), 1701, https://doi.org/10.1038/s41467-021-21655-w, 2021.

Batjes, N. H.: ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2) [data set], https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/82f3d6b0-a045-4fe2-b960-6d05bc1f37c0, 2012~~3~~.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., & Wood, E. F: Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments, Journal of Geophysical Research: Atmospheres, 125(17), https://doi.org/10.1029/2019JD031485, 2020.

Beck, H. E., van Dijk, A. I. J. M., Roo, A. de, Dutra, E., Fink, G., Orth, R. & Schellekens, J.: Global evaluation of runo~~i~~ff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881-20903, https://doi.org/10.5194/hess-21-2881-2017, 2017.

Beck, H. E., van Dijk, A. I. J. M., Roo, A. de, Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resources Research, 52(5), 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

Benjamini, Y., & Hochberg, Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. http://www.jstor.org/stable/2346101, 1995.

Boulange, J, Hanasaki, N, Yamazaki, D., & Pokhrel, Y.: Role of dams in reducing global flood exposure under climate change, Nature Communications, 12(1), 417, https://doi.org/10.1038/s41467-020-20704-0, 2021.

Breimann, L.: Random Forests, Machine Learning, 45, 1–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chaney, N. W., Herman, J. D., Ek, M. B., & Wood, E. F.: Deriving global parameter estimates for the Noah land surface model using FLUXNET and machine learning, Journal of Geophysical Research: Atmospheres, 121(22), 13,218–13,235, https://doi.org/10.1002/2016JD024821, 2016.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, Journal of Statistical Software, 61(6), 1–36. https://doi.org/10.18637/jss.v061.i06, 2014.

Cuntz, M., Mai, J., Samaniego, L, Clark, M., Wulfmeyer, V., Branch, O., Attinger, S, & Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, Journal of Geophysical Research: Atmospheres, 121, 10,676 - 10,700, https://doi.org/10.1002/2016JD025097, 2016.

Döll, P. & Fiedler, K.: Global-scale modeling of groundwater recharge, Hydrol. Earth Syst. Sci., 12, 863–885, https://doi.org/10.5194/hess-12-863-2008, 2008

Döll, P., Kaspar, F., & Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, Journal of Hydrology, 270, 105–13, https://doi.org/10.1016/S0022-1694(02)00283-4, 2003.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., & Kusche, J.: everaging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin, Hydrology and Earth System Sciences, 28 (10), 2259-2295, https://doi.org/10.5194/hess-28-2259-2024, 2024.

Draper, C. S., Walker, J. P., Steinle, P. J., de Jeu, R. A. M., Holmes T. R. H.: An evaluation of AMSR–E derived soil moisture over Australia, Remote Sensing of Environment, 113, 703-710, https://doi.org/10.1016/j.rse.2008.11.011, 2008.

Eisner, S.: Comprehensive Evaluation of the WaterGAP3 Model across Climatic, Physiographic, and Anthropogenic Gradients, Ph.D. thesis, University of Kassel, Kassel, Germany, 128pp., 2016.

Friedl, M., Sulla-Menashe, D.: MCD12Q1 MODIS/Terra+Aqua Land, Cover Type Yearly L3 Global 500m SIN Grid V006, NASA EOSDIS Land Processes DAAC [data set]~~, NASA EOSDIS Land Processes DAAC~~, https://doi.org/10.5067/MODIS/MCD12Q1.006, 2019.

Feigl, M., Thober, S., Schweppe, R., Herrnegger, M., Samaniego, L., & Schulz, K.: Automatic Regionalization of Model Parameters for Hydrological Models, Water Resources Research, 58, e2022WR031966, https://doi.org/10.1029/2022WR031966, 2022.

Golian, S., Murphy, C., & Meresa, H.: Regionalization of hydrological models for flow estimation in ungauged catchments in Ireland, Journal of Hydrology: Regional Studies, 36, 100859, https://doi.org/10.1016/j.ejrh.2021.100859, 2021.

GRDC, The Global Runoff Data Centre, 56068 Koblenz, Germany, 2020.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., & Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe. Journal of Hydrometeorology, 13(2), 604-620. https://doi.org/10.1175/JHM-D-11-083.1, 2012.

Guo Y, Zhang Y, Zhang L, & Wang Z: Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review, WIREs Water, 8, e1487, https://doi.org/10.1002/wat2.1487, 2020.~~1~~

Gupta, H. V, Sorooshian, S., & Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34(4), 751–763, https://doi.org/10.1029/97WR03495, 1998.

He, Y., Bárdossy, A., & Zehe, E.: A review of regionalisation for continuous streamflow simulation, Hydrology and Earth System Sciences, 15(11), 3539–3553. https://doi.org/10.5194/hess-15-3539-2011, 2011.

Jansen, K. F., Teuling, A. J., Craig, J. R., Dal Molin, M., Knoben, W. J. M., Parajka, J., Vis, M., Melsen, L. A.: Mimicry of a conceptual hydrological model (HBV): What's in a name? Water Resources Research, 57, e2020WR029143. https://doi.org/10.1029/2020WR029143, 2022.

Kaspar, F.: Entwicklung und Unsicherheitsanalyse eines globalen hydrologischen Modells, Ph.D. thesis, University of Kassel, Kassel, Germany, 129pp., 2004.

Khosa, F. V., Mateyisi, M. J., van der Merwe, M. R., Feig, G. T., Engelbrecht, F. A., Savage, M. J.: Evaluation of soil moisture from CCAM-CABLE simulation, satellite-based models estimates and satellite observations: a case study of Skukuza and Malopeni flux towers, Hydrology and Earth System Sciences, 24(4), 1587-1609, https://doi.org/10.5194/hess-24-1587-2020, 2020.

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., Olden, J. D.: Assessing placement bias of the global river gauge network, Nature Sustainability, 5, 586–592. https://doi.org/10.1038/s41893-022-00873-0, 2022.

Kupzig, J., Reinecke, R., Pianosi, F., Flörke, M., & Wagener, T.: Towards parameter estimation in global hydrological models, Environmental Research Letters, 18(7), 74023. https://doi.org/10.1088/1748-9326/acdae8, 2023.

Lange, S.: EartH2Observe, WFDEI and ERA-Interim data Merged and Bias-corrected for ISIMIP (EWEMBI), V. 1.1, GFZ Data Services [data set]~~, GFZ Data Services~~, https://doi.org/10.5880/pik.2019.004, 2019.

Lebecherel, L., Andréassian, V., Perrin: On evaluating the robustness of spatial-proximity-based regionalization methods, Journal of Hydrology, 539, 196-203, https://doi.org/10.1016/j.jhydrol.2016.05.031, 2016.

Lehner, B. and Döll, P: Development and validation of a global database of lakes, reservoirs and wetlands, Journal of Hydrology, 296 (1-4), 1-22, https://doi.org/10.1016/j.jhydrol.2004.03.028, 2004.

Lehner, B., Verdin, K., & Jarvis, A.: New global hydrography derived from spaceborne elevation data, Eos, Transactions, AGU, 89, 93–94, doi:10.1029/2008EO100001, 2008.

Liam, A., & Wiener, M.: Classification and Regression by randomForest. R News, 2(3), 18–22, 2002.

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

McIntyre, N, Lee, H., Wheater, H., Young, A., & Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resources Research, 41(12), W12434, https://doi.org/10.1029/2005WR004289, 2005.

Merz, R., Blöschl, G.: Regionalisation of catchment model parameters, Journal of Hydrology, 287, 95-123, https://doi.org/10.1016/j.jhydrol.2003.09.028, 2004.

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat, E., Portmann, F. T., Reinecke, R., Schumacher, M., Shadkam, S., Telteu, C.-E., Trautmann, T., Döll, P.: The global water resources and use model WaterGAP v2.2d: model description and evaluation, Geoscientific Model Development, 14(2), 1037–1079, https://doi.org/10.5194/gmd-14-1037-2021, 2021.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T. A., Schiebener, L., Schumacher, M., Döll, P.: The global water resources and use model WaterGAP v2.2e: description and evaluation of modifications and new features, Geoscientific Model Development Discussions [preprint], 1-46, https://doi.org/10.5194/gmd-2023-213, 2023.

Nijssen, B., O'Donnell, G. M., Lettenmeier, D. P., Lohmann, D., & Wood, E. F.: Predicting the Discharge of Global Rivers, American Meteorological Society, 3307–3323, https://doi.org/10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2, 2000.

Oloruntoba, B., Kollet, S., Motzka, C., Vereecken H., Franssen H.-J. H.: High Resolution Land Surface Modelling over Africa: the role of uncertain soil properties in combination with temporal model resolution, EGUsphere Preprint repository [preprint], https://doi.org/10.5194/egusphere-2023-3132, 2024.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resources Research, 44(3), W03413, https://doi.org/10.1029/2007WR006240, 2008.

Oudin, L., Kay, A., Andréassian, V., & Perrin, C.: Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research, 46(11), W11558, https://doi.org/10.1029/2009WR008887, 2010.

Pagliero, L., Bouraoui, F., Diels, J., Willems, P., & McIntyre, N.: Investigating regionalization techniques for large-scale hydrological modelling, Journal of Hydrology, 570, 220–235, https://doi.org/10.1016/j.jhydrol.2018.12.071, 2019.

Parajka, J., Merz, R., & Blöschl, G.: A comparison of regionalisation methods for catchment model parameters, Hydrology and Earth System Sciences, 9, 157–171, https://doi.org/10.5194/hess-9-157-2005, 2005.

~~Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivaplan, M. & Blöschl, G.: Comparative assessment of prediction in ungauged basins – Part 1: Runoff-hydrograph studies, Hydrology and Earth System Sciences, 17, 1783-1795, www.hydrol-earth-syst-sci.net/17/1783/2013/, 2013.~~

Poissant, D., Arsenault, R. & Brissette, F.: Impact of parameter set dimensionality and calibration procedures on streamflow prediction at ungauged catchments, Journal of Hydrology: Regional Studies, 12, 220–237, https://doi.org/10.1016/j.ejrh.2017.05.005, 2017.

Pool, S., Vis, M., & Seibert, J.: Regionalization for ungauged catchments — Lessons learned from a comparative large-sample study. Water Resources Research, 57, e2021WR030437. https://doi.org/10.1029/2021WR030437, 2021.

Qi, W., Chen, J., Li, L., Xu, C., Li, J., Xiang, Y., & Zhang, S.: A framework to regionalize conceptual model parameters for global hydrological modelling, Hydrology and Earth System Sciences Discussions [preprint], https://doi.org/10.5194/hess-2020-127, 2020.

R Core Team.: R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/, 2020.

Reichl, J. P. C., Western, A. W., McIntyre, N. R. & Chiew, F. H. S: Optimization of a Similarity Measure for Estimating Ungauged Streamflow, Water Resources Research, 45 (10), https://doi.org/10.1029/2008WR007248, 2009

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

Schaefli, B., & Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21(15), 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.

Schweppe, R., Thober, S., Müller, S., Kelbling, M., Kumar, R., Attinger, S., & Samaniego, L.: MPR 1.0: a stand-alone multiscale parameter regionalization tool for improved parameter estimation of land surface models, Geoscientific Model Development, 15, 859–882, https://doi.org/10.5194/gmd-15-859-2022, 2022.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrological Processes, 15(6), 1063–1064, https://doi.org/10.1002/hyp.446, 2001.

Shannon, C. E.: A Mathematical Theory of Communication, The Bell System Technical Journal, 3(27), 379-423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x, 1948.

Stacke, T., & Hagemann, S.: HydroPy (v1.0): a new global hydrological model written in Python, Geoscientific Model Development, 14, 7795–7816, https://doi.org/10.5194/gmd-14-7795-2021, 2021.

Tang, Y., Marshall, L., Sharma, A. & Smith, T.: Tools for investigating the prior distribution in Bayesian hydrology, Journal of Hydrology, 538, 551-562, https://doi.org/10.1016/j.jhydrol.2016.04.032, 2016.

Tongal, H., & Sivakumar, B.: Cross-entropy clustering framework for catchment classification, Journal of Hydrology, 552, 433–446, https://doi.org/10.1016/j.jhydrol.2017.07.005, 2017.

Venables, W. N., & Ripley, B. D.: Modern Applied Statistics with S (Fourth Edition). Springer Science+Business Media New York, USA, 501pp, ISBN 978-1-4419-3008-8, 2002

Wagener, T., Wheater, H. S., & Gupta, H. V.~~ (2004).~~: Rainfall – Runoff Modelling in Gauged and Ungauged Catchments, Imperial College Press, London, UK, 332pp., https://doi.org/10.1142/p335, 2004.

Wagener, T., & Wheater, H. S.: Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, Journal of Hydrology, 320, 132-154, https://doi.org/10.1016/j.jhydrol.2005.07.015, 2006.

Ward, P. J., Jongman, B., Sperna Weiland, F., Bouwman, A., Van Beek, R., Bierkens, M. F. P., Ligtvoet, W., & Winsemius, H. C.: Assessing flood risk at the global scale: model setup, results, and sensitivity, Environmental Research Letters, 8, Article 044019. https://doi.org/10.1088/1748-9326/8/4/044019, 2013

Widén-Nilsson, E., Halldin, S., & Xu, C.: Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation, Journal of Hydrology, 340(1-2), 105–118, https://doi.org/10.1016/j.jhydrol.2007.04.002, 2007.

Wu, H., Zhang, J., Bao, Z., Wang, G., Wang, W., Yang, Y. & Wang, J.: Runoff Modeling in Ungauged Catchments Using Machine Learning Algorithm-Based Model Parameters Regionalization Methodology, Engineering, 28, 93-104, https://doi.org/10.1016/j.eng.2021.12.014, 2023.

Yang, X., Magnusson, J., Huang, S., Beldring, S., & Xu, C.: Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions, Journal of Hydrology, 582, 124357, https://doi.org/10.1016/j.jhydrol.2019.124357, 2020.

Yoshida, T., Hanasaki, N, Nishina, K., Boulange, J, Okada, M., & Troch, P. A.: Inference of Parameters for a Global Hydrological Model: Identifiability and Predictive Uncertainties of Climate-Based Parameters, Water Resources Research, 58, e2021WR03066, https://doi.org/10.1029/2021WR030660, 2022.