https://doi.org/10.5194/gmd-2024-47
Preprint. Discussion started: 18 March 2024

Geoscientific Model Development
Discussions
Open Access

Regionalization 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 is necessary to estimate parameters in ungauged basins. This study presents new regionalization methods for WaterGAP3 and aims to provide insights into selecting a suitable regionalization method and evaluating its impact on the simulation. 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. 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 that regionalization causes spatially varying uncertainty for ungauged regions. For example, India and Indonesia are particularly affected by higher uncertainty. The impact of regionalization in ungauged areas propagates through the water system, e.g., one water balance component changed by approximately 2400 km^3 yr^{-1} on a global scale, which is in the range of inter-model differences. The magnitude of the impact of regionalization 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 modelling 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 is the estimation of parameter values in a model for ungauged basins (Oudin et al., 2008), usually based on information from gauged basins (Oudin et al., 2010). 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 methods exist, and no overall preferred method has been found (Ayzel et al., 2017; Pool et al., 2021). In contrast, the optimal 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 method for a specific use case (e.g., Qi et al., 2020).

Evaluation is needed to assess different regionalization methods. Evaluation is particularly challenging for regionalization methods because they are usually applied when monitoring data is missing. Therefore, 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 at the global level.

In general, 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 is the difficulty of incorporating parameter interdependencies (Poissant et al., 2017). 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. 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 methods with machine learning-based methods for a GHM on a global scale.

Some regionalization methods do not make a clear distinction between calibration and 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 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 method for the state-of-the-art GHM WaterGAP3 (Eisner, 2016). It compares traditional regionalization methods with machine learning-based methods and uses a "benchmark-to-beat" and an ensemble split-sample test to evaluate the applied methods. The overall research topic is evaluating and selecting the most appropriate regionalization method for a GHM. Specifically, the study has two objectives. It aims

(1) to propose a selection for the regionalization method of WaterGAP3 and
(2) to evaluate the impact of an improved regionalization method against a benchmark-to-beat.

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 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. This model parameter is calibrated per basin within the range of 0.1 and 5. The objective function for calibration is to minimize the deviation between the mean annual simulated and observed river discharge. Thus, as a result of the calibration, each basin has a calibrated value (γ) between 0.1 and 5. 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 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, 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. 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 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). 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 standard calibration approach for WaterGAP3. After the standard calibration, some basins still have an insufficient model performance, i.e., more than 20% bias in monthly discharge. These 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 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 basins out of 1,861 basins are selected for regionalization (323 are neglected due to low model performance, and 302 are neglected due to insufficient basin size).

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

Figure 1b shows the calibrated values for γ. It emerges that the calibrated values tend to bet at the upper and lower bounds of the parameter space. This misbehaviour is already known (cf. Fig. 4b in Müller Schmied et al., 2021) and highlights the need to further develop the calibration strategy for WaterGAP3, e.g., by implementing multi-variate calibration. However, this study focuses solely on analysing and implementing a new regionalization method. It does not aim to change the calibration approach of WaterGAP3. To achieve the latter, future studies are

https://doi.org/10.5194/gmd-2024-47
Preprint. Discussion started: 18 March 2024

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).

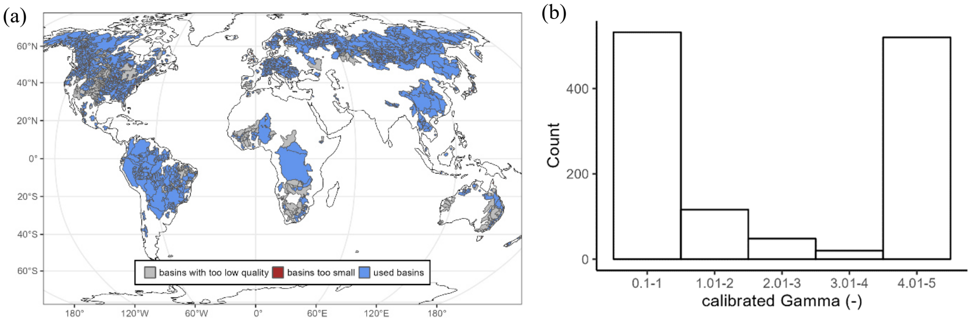

Figure 1: (a) Gauged basins calibrated beforehand, highlighting basins not used for regionalization due to low model performance or too small basin size and (b) the histogram of the calibrated model parameter values of all used basins showing heavy-tails.

2.3 Basin Descriptors

This study uses basin descriptors as predictors to drive regression-based or distance-based regionalization approaches. These basin descriptors are based on model data and are aggregated to basin values using a simple mean method to have the exact 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 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. In this study, we use 12 basin descriptors to develop regionalization methods; nine of these descriptors are physiographic, while the remaining three are climatic (see Table 1). Most descriptors are not correlated (see Appendix A), i.e., we avoid redundant information (Wagener et al., 2004).

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.

The correlation coefficients and the corresponding information gain are listed in Table 1. All basin descriptors have a low correlation coefficient, e.g., the highest Pearson correlation is -0.36. 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, with a maximum of 14.4% of the information explained by the temperature descriptor. 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, the calibrated value masks the effect of multiple sources of errors, such as uncertainty in the input data, model structure, or varying hydrological processes. 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. We implement regionalization methods using four groups of basin descriptors by selecting basin descriptors with the highest correlation coefficients and information gain:

- “cl”: two correlated climatic descriptors (mean temperature, annual shortwave radiation),
- “p”: three correlated physiographic descriptors (slope class, forest %, permafrost %),
- “p+cl”: two correlated climatic & three physiographic descriptors, 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 Methods

In our study, we test several traditional and machine learning-based regionalization methods against each other and a defined benchmark-to-beat to find the most suitable regionalization method for WaterGAP3. At the global scale, 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 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: 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.

To evaluate the regionalization methods, we implement an ensemble of split-sample tests. Specifically, we randomly split the basins into 50% gauged and 50% pseudo-ungauged basins. This 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 randomly selected 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 and calibrated γ. Thus, the mean absolute error (MAE), an easy-to-interpret measure, is used to evaluate the prediction accuracy. The lower the MAE, the better the prediction; an MAE of zero 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 method is robust if the prediction accuracy is similar in training and testing. A generally good performance, i.e., small MAE values, indicates that the regionalization method suits WaterGAP3.

Regression-based methods

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. 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 that the algorithm can choose between two randomly selected predictors at each node. We use 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 a traditional physical similarity approach, we use Similarity Indices (in the following named with SI). We use the methodology proposed by Beck et al. (2016). The SI (see Eq. (2)) are derived using the basin descriptors 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 “10”), which is based on Beck et al. (2016) and McIntyre et al. (2006). Further, we apply a simple mean method for the ensemble-based prediction to aggregate the ensemble values into one predicted parameter value.

https://doi.org/10.5194/gmd-2024-47
Preprint. Discussion started: 18 March 2024

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

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 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 variation between groups and minimize variation within groups. We 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 (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 is less flexible than SI, we implement a highly flexible version of k-means with 162 groups, where each ungauged basin is sorted into a very small basin group. Using this highly flexible version of k-means, we test whether the potential differences between SI and k-means are based on the degree of flexibility.

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

where $Z'_{p,j}$ is the normalised 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.

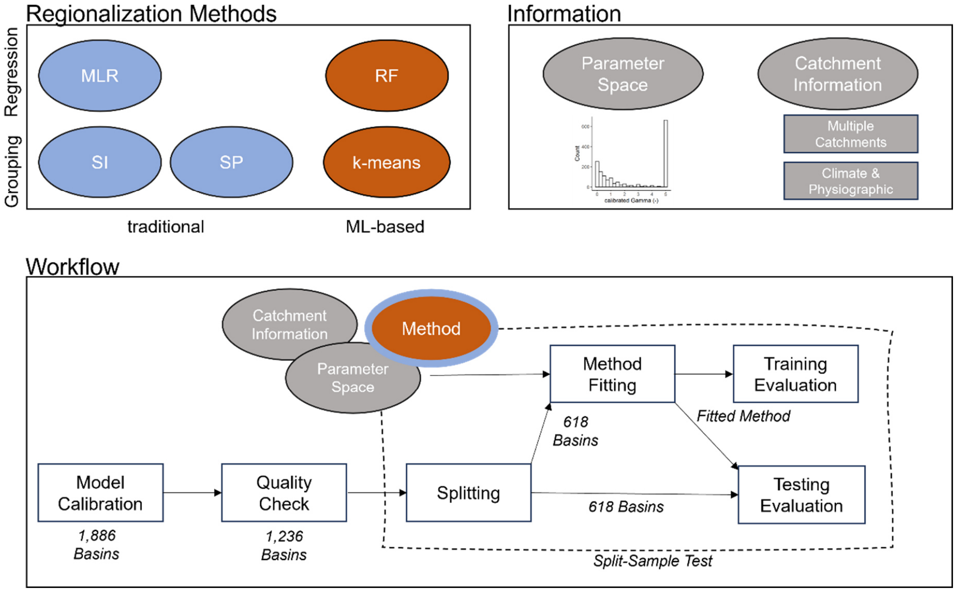

Figure 2: Experimental setup of the study: regionalization methods, used modifications and information and the general workflow (MLR: Multiple Linear Regression, SI: Similarity Indices, SP: Spatial Proximity, RF: RandomForest).

Spatial Proximity

The spatial proximity approach is one of the easiest to regionalize parameter values. However, it is also often criticized that nearby basins do not necessarily have the same hydrological behaviour (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 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 provides an overview of the applied regionalization methods and information used for the experimental setup.

3. Results and Discussion

3.1 Evaluating Traditional Methods

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). 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.

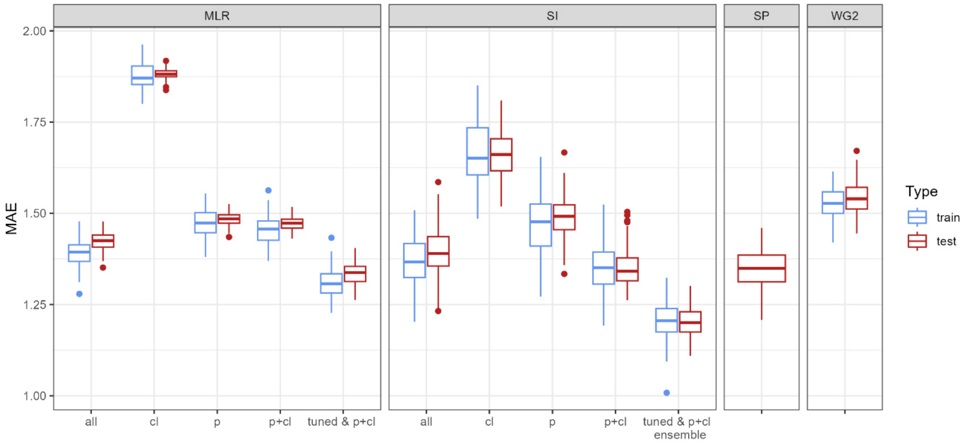

Figure 3: Split-sampling results for the benchmark-to-beat taken from WaterGAP2 (WG2) and different versions of the traditional regionalization methods: Multiple Linear Regression (MLR), Similarity Indices (SI) and Spatial Proximity (SP).

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

https://doi.org/10.5194/gmd-2024-47
Preprint. Discussion started: 18 March 2024

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).

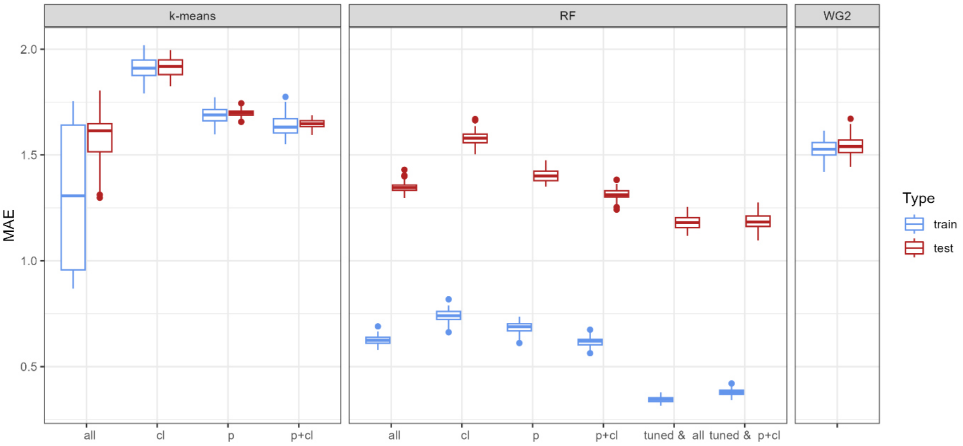

Figure 4: Split-sampling results for the benchmark-to-beat taken from WaterGAP2 (WG2) and different versions of machine learning-based approaches: k-means (in combination with knn) and RandomForest (RF).

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.

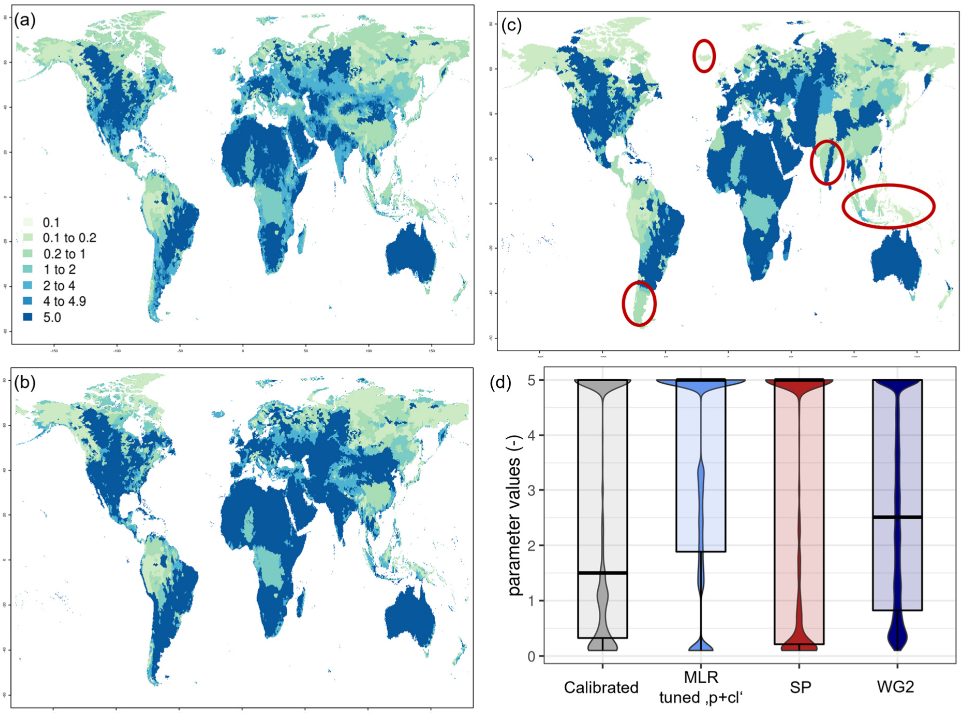

Figure 5: Global γ distribution for different regionalization methods, highlighting areas of differences (a) γ distribution using the MLR approach with parameter space tuning, using physiographic and climatic basin descriptors as independent variables, i.e., tuned MLR "p+cl", (b) benchmark-to-beat, WG2,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

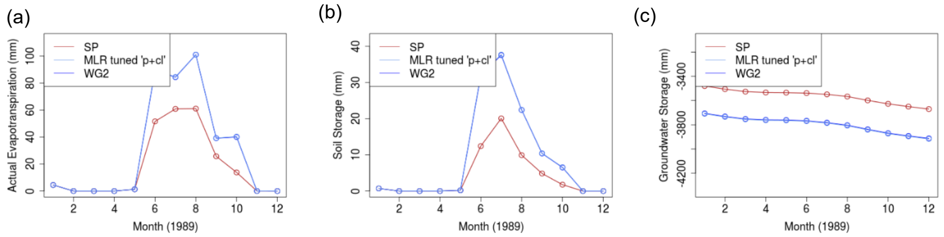

Figure 6: Differences in monthly internal states and fluxes of WaterGAP3 for one grid cell with varying regionalized value (SP: 0.325, MLR tuned "p+cl": 5 and benchmark-to-beat (WG2): 4.467243), located in India (21.519794°|70.566733°) for a) actual evapotranspiration, b) soil storage and c) groundwater 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.

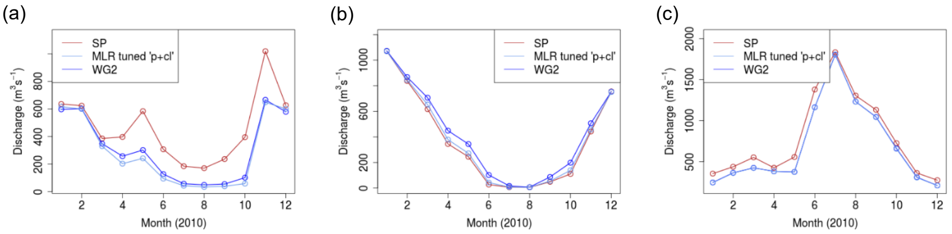

Figure 7: Simulated monthly runoff using three different regionalization methods for a) the Tiber, b) the Ebro and c) Rio Negro (in Argentina) for 2010 as an exemplary year.

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 km³ 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 km³ 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.

4. 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 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.

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 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.

https://doi.org/10.5194/gmd-2024-47
Preprint. Discussion started: 18 March 2024

Appendix 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. The information is the product of rooting depth (defined in a look-up table) and the total available water content derived from Batjes (2013).
- *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 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.
- *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 is shown in Fig. A2.

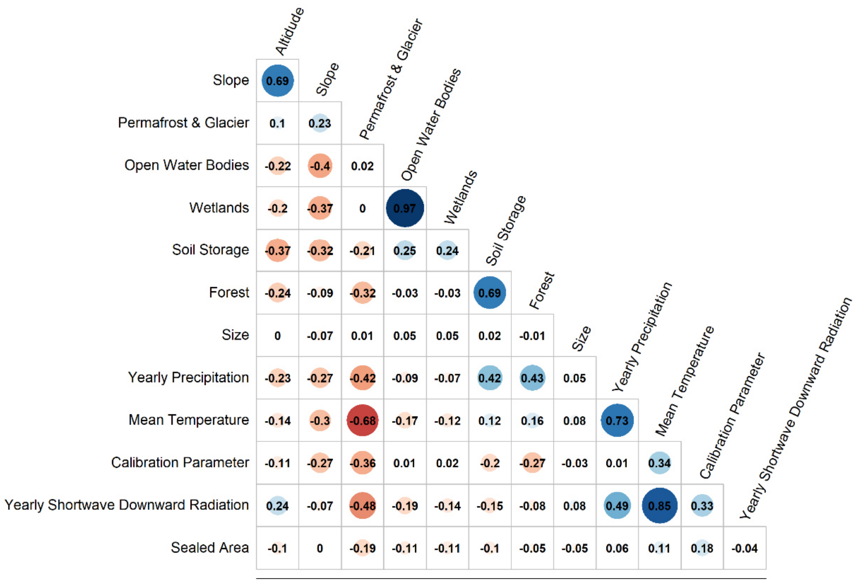

Figure A1: Correlation between basins descriptors.

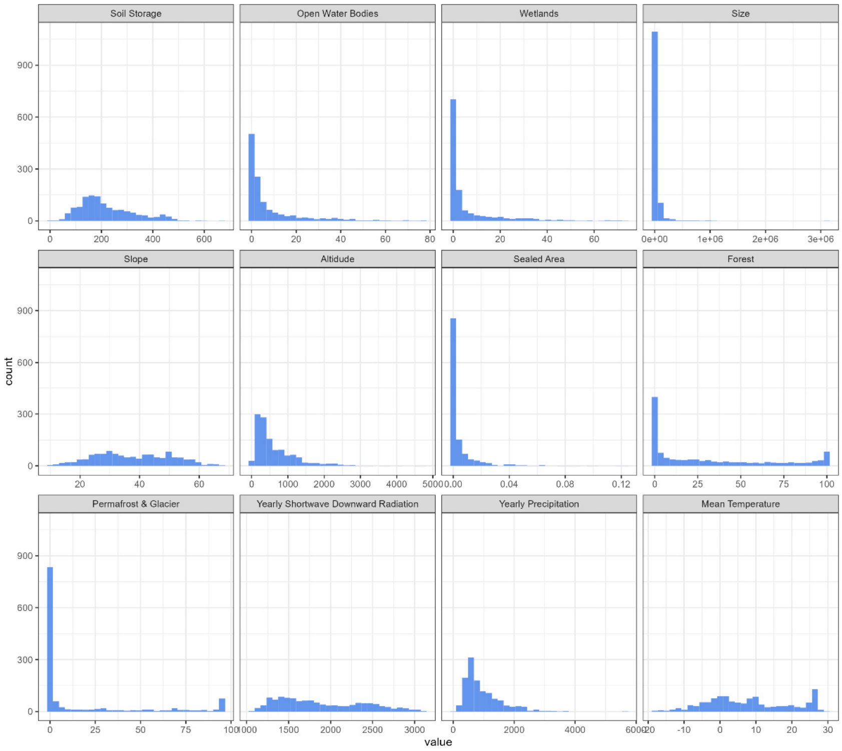

Figure A2: Distribution of basins descriptors within all basins used for regionalization (n=1,236)

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

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, 2013.

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 runiff 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.

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.

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.

Eisner, S.: Comprehensive Evaluation of the WaterGAP3 Model across Climatic, Physiographic, and Anthropogenic Gradients, PhD 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 [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.

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, 2021

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.

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

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 [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.

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.

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.

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.

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.