AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research®

RESEARCH ARTICLE
10.1029/2021WR030437

Regionalization for Ungauged Catchments — Lessons Learned From a Comparative Large-Sample Study

Sandra Pool[1,2], **Marc Vis**[3], **and Jan Seibert**[3,4]

[1]Department Water Resources and Drinking Water, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [2]Department Systems Analysis, Integrated Assessment and Modelling, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [3]Department of Geography, University of Zurich, Zurich, Switzerland, [4]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Key Points:

- A comparison of 19 regionalization approaches was conducted using 671 U.S. catchments and a homogenized modeling protocol
- Almost perfect donors exist, and excellent relative performance can be reached for most catchments with current regionalization approaches
- The ranking of regionalization approaches depends more on how the predicted hydrographs are evaluated than how the donors are calibrated

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Pool,
sandra.pool@eawag.ch

Citation:

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

Received 19 MAY 2021
Accepted 19 AUG 2021

Abstract Model parameter values for ungauged catchments can be regionalized from hydrologically similar gauged catchments. Achieving reliable and robust predictions in ungauged catchments by regionalization, however, is still a major challenge. Here, we conduct a comparative assessment of 19 regionalization approaches based on previously published literature to contribute new insights into their performance in different geographic regions. The approaches use geographical information, physical catchment attributes, hydrological signatures, or a combination thereof to select donor catchments and to subsequently transfer their entire parameter sets to the ungauged receiver catchment. Each regionalization approach was tested in a leave-one-out cross-validation with a bucket-type catchment model (the HBV model) using 671 gauged catchments in the United States with a diverse hydroclimatology. We then evaluated regionalization performance for several hydrograph aspects, compared it against calibration and regionalization benchmarks, and linked it to catchment descriptors. The results of this large-sample regionalization study can be summarized in three major lessons: (a) Catchments can benefit from a well-chosen regionalization approach independent of their geographic region and independent of how well they can be modeled or regionalized at best. (b) Almost perfect donors exist for most catchments and an excellent relative model performance can be reached for most catchments with current regionalization approaches. This implies that there is considerable potential for improvement in the prediction in ungauged catchments. (c) The ranking of regionalization approaches depends on how the predicted hydrographs are evaluated. These findings indicate that a multi-criteria evaluation is essential for a robust assessment of regionalization performance.

Plain Language Summary Information on streamflow is crucial for good water resources management including the mitigation of water-related hazards. However, for many catchments there is a lack of streamflow information. In such situations, streamflow is often estimated using hydrological models, whereby model parameterizations are transferred (i.e., regionalized) from hydrologically similar gauged catchments. Reliable estimates in data-scarce regions are still a major challenge in hydrology despite the large number of regionalization approaches proposed in the past decades. Here, we conduct a systematic and standardized assessment of 19 existing regionalization approaches using 671 catchments in the United States. Our findings suggest that widely used regionalization approaches can result in excellent model performance for most catchments, whereby approaches considering spatial proximity and any kind of volume information are among the most promising ones. While volume information is per definition missing in ungauged catchments, it could possibly be derived from a small number of field measurements or estimated through statistical analysis. However, the most suitable approach can vary considerably among catchments, and an improved understanding of the characteristics and parameter values of great donors and their relationship to an ungauged catchment will be key to advance regionalization further.

1. Introduction

1.1. Regionalization of Hydrological Model Parameter Values

Continuous daily streamflow information is crucial for many practical purposes related to assessing and managing water resources or water-related hazards. However, many catchments of interest are poorly

gauged or ungauged and lack the required hydrological information. Streamflow predictions in ungauged catchments are traditionally based on hydrological models (Parajka et al., 2013), whereby model parameter values in the ungauged catchment need to be inferred (regionalized) based on information from other gauged catchments (Blöschl & Sivapalan, 1995). Achieving robust and reliable predictions in ungauged catchments by regionalizing model parameters is a major challenge in hydrology (Hrachowitz et al., 2013).

Many approaches have been proposed to regionalize model parameters (see reviews by, e.g., He et al., 2011; Hrachowitz et al., 2013; Parajka et al., 2013; Razavi & Coulibaly, 2013). The most widely applied regionalization approaches can be broadly classified into two groups: (a) regression-based methods that relate individual model parameters to catchment characteristics (e.g., Seibert, 1999; Skaugen et al., 2015; Song et al., 2019) and (b) distance-based methods where entire parameter sets are transferred between hydrologically similar catchments using spatial proximity or catchment attributes (see early work by, e.g., Kokkonen et al., 2003; McIntyre et al., 2005; Oudin et al., 2008; Parajka et al., 2005), or any hydrological information available (e.g., Masih et al., 2010; Pool et al., 2019; Rojas-Serna et al., 2016) as a similarity metric. Regression-based approaches have been criticized for ignoring equifinality and the dependency of values for the different model parameters (Arsenault & Brissette, 2014; Bárdossy, 2007; McIntyre et al., 2005). Consequently, correlations between parameter values and catchment attributes are often weak or cannot be hydrologically justified (Oudin et al., 2008; Seibert, 1999; Skaugen et al., 2015). Distance-based approaches acknowledge parameter dependency by transferring entire parameter value sets; however, their success depends on identifying a suitable similarity metric. For example, spatial proximity may only be an appropriate similarity metric in data-rich regions (Lebecherel et al., 2016; Neri et al., 2020) or where hydrological processes vary smoothly in space (He et al., 2011). On the other hand, the use of catchment attributes for defining similarity can be confounded by the lack of characteristics representing hydrological processes (Oudin et al., 2010) or missing spatial patterns of characteristics controlling model parameter values (Merz et al., 2020).

While all these regionalization approaches have their strengths and weaknesses, there is a tendency toward a somewhat lower performance of regression-based methods than distance-based methods (Arsenault & Brissette, 2014; Bao et al., 2012; McIntyre et al., 2005; Oudin et al., 2008; Yang et al., 2018). However, blending regression with spatial proximity or attribute similarity (Arsenault & Brissette, 2014), combining spatial proximity and attribute similarity (Zhang & Chiew, 2009; Yang et al., 2018), or considering (regionalized) streamflow information (Masih et al., 2010; Pool et al., 2019; Rojas-Serna et al., 2016) can improve predictions in ungauged catchments. The review of Parajka et al. (2013) further highlights the important role of climate for regionalization performance. While the performance is generally higher in humid catchments than in arid catchments, the most successful regionalization approach is likely different for a humid catchment than an arid catchment. Yang et al. (2020) showed that climate may outweigh the difference between regionalization approaches even within a relatively narrow range of humid high-latitude climates. Despite these general tendencies, there is considerable disagreement among studies regarding choosing the most appropriate regionalization approach (He et al., 2011; Parajka et al., 2013; Razavi & Coulibaly, 2013). Indeed, Parajka et al. (2013) found that the performance for a given approach typically differs more between studies than between approaches tested within a single study.

One reason for the lack of consensus among regionalization studies could be related to the use of different performance metrics. The choices related to these metrics can affect regionalization performance in multiple ways. First, the performance metric used during the calibration process directly affects model parameter values (see review by, e.g., Efstratiadis & Koutsoyiannis, 2010). The calibrated parameter values of a gauged catchment are the foundation of regionalization and can significantly influence donor suitability (Singh et al., 2014) or regression models (Song et al., 2019). Calibration metrics have so far received little attention in regionalization studies, and their impact on the success of a given regionalization approach has not yet been explored in a comparative large-sample study. Second, performance metrics largely define which hydrograph aspects and catchment processes a simulation is evaluated on (e.g., Legates & McCabe, 1999; Madsen, 2000; Yilmaz et al., 2008). The separate consideration of several hydrograph aspects enables a comprehensive assessment of regionalization performance (e.g., Viglione et al., 2013; Rojas-Serna et al., 2016; Yang et al., 2020) and facilitates their comparison between studies (Parajka et al., 2013). Despite the benefits of such multi-objective evaluations, these are still not consistently applied. Finally, the success of regionalization approaches is often judged based on the value of the chosen performance metric. However, the same

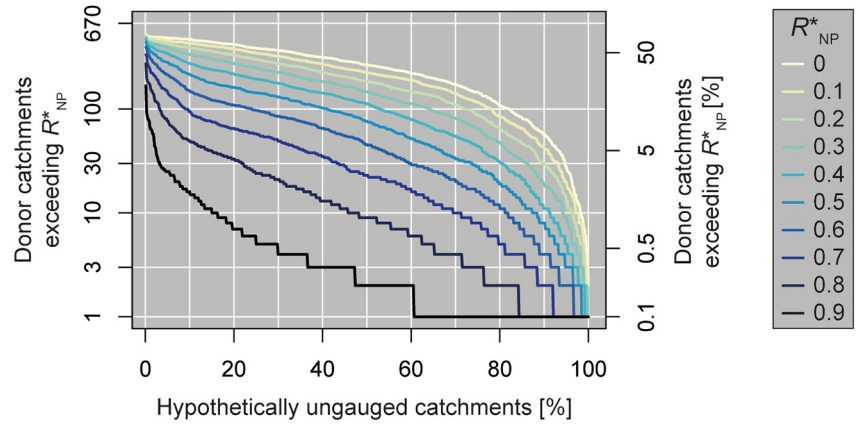

Figure 1. Opportunities and challenges of finding suitable donors in a hydrologically diverse large-sample data set with 671 catchments in the United States. Each of the 671 catchments was calibrated using R_{NP} (see Section 2.2.2 for more details), and the resulting parameter sets were evaluated for their performance as a donor for each of the other 670 catchments. Calibration and evaluation were conducted during the simulation period from October 1, 1989 to September 30, 1999. The lines indicate the percentage of donor catchments giving a particular regionalization performance for a hypothetically ungauged catchment. The performance R^*_{NP} equals 0 for the regionalization with random parameter values and R^*_{NP} equals 1 for a local calibration of the hypothetically ungauged catchment.

performance value can have a different meaning for different catchments. A value might be considered to have a good performance in one catchment (e.g., if hydrological processes are complex or observation uncertainty is high), while the same value might indicate a rather unsatisfactory performance in another catchment (e.g., if the hydrological response is strongly linked to precipitation or high-quality data are available). This complicates the comparison of model performance across catchments, in particular, if they have contrasting runoff regimes (Schaefli & Gupta, 2007; Seibert et al., 2018). Therefore, Seibert et al. (2018) advocated the use of relative model performance metrics that are based on an upper and a lower benchmark which are computed from simulations based on a local calibration and a random parameterization. In the context of regionalization, we suggest extending the concept of upper and lower benchmarks with regionalization benchmarks.

The comparison of regionalization studies is further complicated by the varying diversity and number of study catchments. A considerable number of large-sample studies (including several tens to thousands of catchments) were conducted with relatively humid catchments of varying snowiness in Europe (see recent work by, e.g., de Lavenne et al., 2019; Neri et al., 2020; Merz et al., 2020; Yang et al., 2020) and Canada (e.g., Arsenault & Brissette, 2014; Razavi & Coulibaly, 2016). In contrast, studies including (semi-) arid catchments are less abundant and were often based on a smaller number of catchments (about 20 to hundred catchments; Bao et al., 2012; Petheram et al., 2012; Post, 2009). The more recently published CAMELS large-sample data set with 671 catchments in the United States (Addor et al., 2017; Newman et al., 2015) may bring new insights into the spatially varying value of regionalization approaches. The large range of streamflow responses encountered in the CAMELS data set can further enhance the chance of finding hydrologically similar donor catchments and potentially reduce the risk of isolated catchments (Bárdossy, 2007; Oudin et al., 2010).

1.2. An Example of Opportunities and Challenges of Finding Suitable Donors in a Large-Sample Data Set

A fundamental question when using the donor catchment approach is whether there are suitable donors at all and how common these are. An initial analysis indicated opportunities and challenges of finding suitable donor catchments in the CAMELS data set. We calibrated the HBV model to all 671 catchments and, for each catchment, tested the other 670 catchments as a donor for the parameter values. Figure 1 shows the percentage of donor catchments exceeding certain relative performance values (R^*_{NP}) as defined in Seibert et al. (2018). A relative performance value of zero corresponds to a donor catchment performing

as well as randomly selected parameter sets. In contrast, a value of one corresponds to a donor catchment with a similar performance as calibrating the receiver catchment. From the line corresponding to a relative performance of zero, one can see that for 50% of the hypothetically ungauged catchments, more than 40% of the donor catchments perform at least as well as randomly selected parameter sets. For the other 60% of the donor catchments, one would be better off using random parameter sets. The set of potential donor catchments becomes significantly smaller for higher relative performances. For a relative performance value of 0.7, half of the hypothetically ungauged catchments have at most 23 suitable donor catchments. However, at least three potential donors exist for 85% of the hypothetically ungauged catchments. When increasing the relative performance threshold further to 0.9, there is at least still one suitable donor catchment for each of the hypothetically ungauged catchments. On the other hand, for 40% of the catchments, only one donor exceeds the performance threshold. In other words, although sometimes limited in number, potentially suitable donor catchments exist, but the challenge remains to find these suitable donors. What are their characteristics, and to which extent are the different regionalization approaches able to find these best donor catchment(s)?

1.3. Scope of This Study

Following the initial analysis presented in the previous section, we conducted a comparative assessment of 19 distance-based regionalization approaches using the hydrologically diverse CAMELS data set with 671 catchments in the contiguous United States. The approaches were defined based on previously published literature, whereby geographical information, physical catchment attributes, hydrological information, or any combination thereof was used to select donor catchments. The performance of the tested regionalization approaches was evaluated in a leave-one-out cross-validation approach, and using the semi-distributed HBV-light model (Seibert & Vis, 2012). With the comparative assessment, we aim to contribute new insights into the performance of different regionalization approaches through a systematic comparison and a homogenized modeling protocol (i.e., identical model structure, as well as calibration and evaluation processes). In this study, we address three long-standing research questions from a new or extended perspective:

1. What performance can be expected for predictions in different geographic regions?
   We address this question by introducing the idea of a regionalization benchmark. This benchmark provides a realistic reference of what could be achieved at best and what should be expected at least with regionalized parameter sets in different geographic regions.
2. Is there a best regionalization approach?
   We complement the common practice of ranking regionalization approaches with an assessment of the robustness of such rankings when using different performance metrics. The ranking at the continental scale is accompanied by a catchment-specific search for the best regionalization approach.
3. What makes a good donor catchment?
   With this question, we move beyond the classical search for the best regionalization approach and instead start exploring the characteristics and parameter sensitivity of the best available donor catchments.

2. Data and Methods

2.1. Study Catchments

This study is based on 671 catchments in the contiguous United States with minimal human influence (Newman et al., 2015). The catchments are distributed across the major watersheds of the United States (HUC2 region; USGS, 2020; Figure A1) and cover a wide range of climatic conditions, topographic aspects, and surface and subsurface properties (Table 1; Addor et al., 2017). The climatic conditions dominate the runoff response at the continental scale (Berghuijs et al., 2014; Jehn et al., 2020), which leads to five major runoff regimes (Brunner et al., 2020): (a) a New Year's regime in the Northwest, (b) a snowmelt dominated regime in the Rocky Mountains, (c) an intermittent runoff regime in the central part of the United States, (d) a weak winter regime along the Atlantic Coast and in the Great Lakes region, and (e) a strong winter regime in the Appalachian region of the eastern United States. The runoff response within these regimes

Table 1
Summary Statistics of Catchment Characteristics of the 671 Study Catchments

Physical attributes and hydrological signatures	5th percentile	Median	Mean	95th percentile
Area [km^2]	22	341	808	2,921
Aridity[a] [–]	0.36	0.86	1.06	2.37
Precipitation seasonality[b] [–]	−1.14	0.08	−0.04	0.74
Snowfall fraction [–]	0.00	0.10	0.18	0.67
Wetland fraction [–]	0.00	0.00	0.16	0.95
Forest fraction [–]	0.00	0.81	0.64	1.00
Clay faction [–]	0.06	0.19	0.20	0.37
Runoff ratio[c] [–]	0.05	0.35	0.39	0.85
Mean daily discharge [mm/day]	0.07	1.13	1.49	5.38
Low flows (Q5) [mm/day]	0.00	0.08	0.17	0.65
High flows (Q95) [mm/day]	0.22	3.77	5.06	16.52
Mean half-flow date [Julian date]	135	174	183	246
Recession slope [mm/day]	0.04	0.09	0.11	0.20

Note. The catchment characteristics are divided into physical attributes and hydrological signatures. Histograms for the physical attributes and hydrological signatures are shown in Figure S1.
[a]Aridity is defined as the ratio of potential evaporation to precipitation. [b]Precipitation seasonality, whereby positive (negative) values indicate a tendency for summer (winter) precipitation. [c]Runoff ratio is defined as the ratio of discharge to precipitation.

is relatively diverse in the western regions, where topographic aspects and subsurface properties can vary considerably over short distances (Jehn et al., 2020).

Data for each catchment were retrieved from the CAMELS data set that provides daily hydrometeorological time series (Newman et al., 2015), and a large number of catchment attributes (Addor et al., 2017) aggregated at the catchment scale. The meteorological time series for each catchment were derived from the spatially distributed Daymet data set (1 km by 1 km) by calculating area-weighted catchment mean values (Newman et al., 2015). Here, the meteorological data of Newman et al. (2015) was further used to calculate monthly potential evaporation with the Priestley-Taylor equation (Priestley & Taylor, 1972). Additionally, we extracted the wetland fraction of each catchment from the Global Lakes and Wetlands Database (Lehner & Döll, 2004), and computed the catchment average recession slopes using the *EflowStats* R-package (USGS, 2014). The computed wetland fraction and recession slopes are provided in the Data Set S1. Detailed topographic information was obtained from the SRTM data (90 m by 90 m; Jarvis et al., 2008).

2.2. Rainfall-Runoff Model Structure and Calibration

2.2.1. The HBV Model

The regionalization approaches were used to transfer the model parameters of the HBV rainfall-runoff model (Bergström, 1976; Lindström et al., 1997). HBV is a bucket-type model, with four routines and 12 parameters that simulate the hydrological response to daily temperature, daily precipitation, and long-term mean monthly potential evaporation. Temperature and precipitation are first input to the snow routine in which snow accumulation and melt are computed using a degree-day method. Snowmelt and rainfall infiltrate into the soil routine from which actual evaporation and groundwater recharge occur as a function of soil water content and potential evaporation. The groundwater routine consists of a shallow storage with two outflows and a deep storage with one outflow that generate peak flow, intermediate flow, and baseflow, respectively. All three flow components are summed up and in the routing routine and transformed by a triangular weighting function to simulate discharge at the catchment outlet. More details on the model structure and parameters can be found in Seibert and Vis (2012) and the supporting information (Figure S2 and Table S1).

In this study, we used the semi-distributed HBV-light version (Seibert & Vis, 2012), whereby each catchment was divided into elevation bands of 200 m. The semi-distributed character simulates snow dynamics and soil-water-related processes for each elevation band separately while keeping a single groundwater storage for the entire catchment. To create the semi-distributed forcing input, we linearly interpolated the catchment mean temperature and precipitation data of the CAMELS data set from the area-weighted reference elevation to all elevation bands using a constant lapse rate of −0.6°C per 100 m (Wallace & Hobbs, 2006) and 10% per 100 m (Johansson, 2000), respectively. The methodology ensures that the catchment mean values of temperature and precipitation are unchanged compared to the Dayment data while representing an elevation effect on temperature and precipitation. For potential evaporation, we used a constant areal mean value within the entire catchment area.

2.2.2. Model Calibration

The HBV-light model was calibrated for each catchment using continuous daily discharge and meteorological data between October 1, 1989 and September 30, 1999. The two years preceding the calibration period were used for model warming-up to start simulations from realistic initial storage values. Model parameter values were optimized within predefined feasible ranges (adapted from Seibert & Vis, 2012) using a genetic algorithm (Seibert, 2000) that selected and recombined a randomly created initial population with 50 parameter sets over 3,500 simulation runs. The optimization of parameter values was repeated 10 times to account for parameter uncertainty and equifinality (Beven & Freer, 2001).

We conducted two independent model calibrations using the Kling-Gupta efficiency R_{KG} (Gupta et al., 2009) and its non-parametric variant R_{NP} (Pool et al., 2018) as objective functions. Both objective functions comprise three error terms that are combined into a scalar using the Euclidean distance. The three error terms evaluate (a) mean discharge using the bias in mean discharge β, (b) flow variability using the bias in the standard deviation α_{KG} for R_{KG} and the absolute error in the normalized flow duration curve α_{NP} for R_{NP}, and (c) flow dynamics using the Pearson correlation r_P for R_{KG} and the Spearman rank correlation r_S for R_{NP}. Differences in the formulation of the variability and dynamic terms of R_{KG} and R_{NP} were shown to be reflected in the simulated hydrographs. R_{KG} tends to focus on the magnitude and timing of high flows, whereas R_{NP} leads to a more balanced evaluation of a broad range of hydrograph aspects (Pool et al., 2018). The mathematical formulations of R_{KG} (Equation 1; Gupta et al., 2009) and R_{NP} (Equation 2; Pool et al., 2018) are as follows:

$$R_{KG} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{KG} - 1)^2 + (r_P - 1)^2} \quad (1)$$

$$R_{NP} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{NP} - 1)^2 + (r_S - 1)^2} \quad (2)$$

with: $\beta = \overline{Q_{sim}}/\overline{Q_{obs}}$; $\alpha_{KG} = \sigma_{Q_{sim}}/\sigma_{Q_{obs}}$; $\alpha_{NP} = 1 - 1/2 \sum_{k=1}^{n} \left| Q_{sim}(I(k))/n\overline{Q_{sim}} - Q_{obs}(J(k))/n\overline{Q_{obs}} \right|$; r_P = Pearson correlation coefficient; r_S = Spearman rank correlation coefficient. Q_{sim} and Q_{obs} are simulated and observed discharge, $Q_{sim}(I(k))$ and $Q_{obs}(J(k))$ are simulated and observed discharge with rank k, n is the length of the time series, and σ is the standard deviation.

2.3. Regionalization

2.3.1. Regionalization Approaches

Many regionalization approaches have been proposed over the last decades for transferring entire parameter sets from gauged donor catchments to an ungauged receiver catchment (see reviews by, e.g., He et al., 2011; Parajka et al., 2013; Razavi & Coulibaly, 2013). In this study, we assessed the performance of 19 regionalization approaches that we defined based on previously published hydrological literature. The regionalization approaches follow two main strategies to select donor catchments: (a) Five approaches select donors from a prior catchment classification (Gottschalk et al., 1979). We chose five classifications representing a wide range of classification variables to ensure diversity in the selection of donors for regionalization. The tested classifications are either based on geographical aspects (USGS, 2020), catchment attributes (Berghuijs et al., 2014), or hydrological signatures (Brunner et al., 2020; Jehn et al., 2020; Schaller & Fan, 2009) and consist of a varying number of classes and class members. For regionalization, all gauged

catchments within the same class as the ungauged catchment were selected as donor catchments. (a) 14 approaches select donors from flexibly defined regions of similar catchments that are delineated specifically for every ungauged catchment (i.e., region of influence approach proposed by Burn, 1990). Catchment similarity was thereby defined using the Euclidean distance (Burn, 1990; McIntyre et al., 2005) in the geographical space, attribute space or signature space. Since these catchment descriptors have different units and distributions, they were standardized with a z-transformation before calculating the similarity metric (Milligan & Cooper, 1988). The three gauged catchments most similar to the ungauged catchment were finally selected as donors.

The suite of regionalization approaches evaluated in this study was complemented by four benchmark methods leading to a total of 23 methods to be tested for each study catchment. We grouped the tested methods into six major categories based on the type of information they use for the selection of donor catchments (for a description of the approaches, we refer to Table 2):

A) Benchmark methods: Upper and lower benchmarks (*UB*, *Best*, *LB*, *and US 670*) were defined to evaluate what could be achieved at best for a given catchment if there was complete information and what should be expected without any prior information (Seibert et al., 2018).
B) Methods that do not include volume or distance information: Donor selection is based on catchment attributes (*Attr*; e.g., McIntyre et al., 2005; Zhang & Chiew, 2009; Oudin et al., 2010), climate classification (*Climate class*; Berghuijs et al., 2014), or a random choice (*Random*; e.g., Zhang & Chiew, 2009).
C) Methods that include distance information: Geographical information is the only criteria for selecting donor catchments in the spatial-proximity approach (*Dist*; e.g., Lebecherel et al., 2016; Neri et al., 2020; Oudin et al., 2008; Parajka et al., 2005), the catchment classification (*Geogr class*; USGS, 2020), and the direct transfer of the simulated hydrographs (*Qsim transfer*; e.g., Seibert, 1999). Spatial proximity can also be combined with catchment attributes to select donor catchments (*Dist & Attr*; e.g., Zhang & Chiew, 2009; Yang et al., 2018).
D) Methods that include volume information: Here, it is assumed that it is possible to estimate hydrological signatures for the ungauged catchment (e.g., Masih et al., 2010) or to directly conduct a number of discharge measurements. Such hydrological information can be used to select donors from a water balance or a signature classification (*WB class*, *Regime class*, and *Sign class*; Brunner et al., 2020; Jehn et al., 2020; Schaller & Fan, 2009), to choose donors based on the similarity in the signature space (*Sign*; e.g., Masih et al., 2010) or the combined signature and attribute space (*Attr & Sign*), or to determine donors based on their ability to reproduce measured streamflow aspects (*Vol* and *RMSE*; e.g., Pool et al., 2019; Rojas-Serna et al., 2016; Viviroli & Seibert, 2015).
E) Methods that include distance and volume information: The approaches tested here consider spatial proximity and signature similarity (combined with attribute similarity; *Dist & Vol*, *Dist & Sign*, and *Dist & Attr & Sign*) or use spatial proximity and discharge measurements to select donor catchments (*Dist & RMSE*; e.g., Pool et al., 2019; Rojas-Serna et al., 2016). These approaches also require either estimates of hydrological signatures or a small number of local point observations.
F) Methods without the transfer of parameter values: Finally, we evaluated the prediction performance when directly transferring the observed hydrographs from the spatially closest catchments (*Qobs transfer*; e.g., Parajka et al., 2015; Patil & Stieglitz, 2012; Razavi & Coulibaly, 2016). We considered this approach as a separate category as it does not include a hydrological model.

Note that some of the tested approaches share one or more criteria to select donor catchments, leading to the selection of similar donor catchments. The percentage of common donors for all combinations of regionalization approaches is shown in Figure S3.

2.3.2. Performance Evaluation

The regionalization approaches were applied in a leave-one-out cross-validation, where each catchment was treated as ungauged at a time, and its streamflow was simulated with the information of the donor catchments. More specifically, each donor catchment provided its 10 calibrated parameter sets to the ungauged catchment. This resulted in 30 model parameterizations if donors were selected from flexibly defined similarity regions or 40 to 2,300 model parameterizations if the selection of donors was based on a prior catchment classification. All parameterizations were used to simulate streamflow in the ungauged

Table 2
Description of the 18 Regionalization and 4 Benchmark Approaches Tested in This Study

Approach		Source of parameter values
Benchmark methods		
UB	–	The upper benchmark corresponds to the local calibration of a catchment.
Best	Flexible	The three best donor catchments (defined by R_{NP}) available from the pool of all 670 potential donors are used as donors.
LB	–	The lower benchmark consists of 1,000 randomly selected parameter values.
US 670	Flexible	All 670 potential donor catchments are used as donors.
Methods that do not include volume nor distance information		
Climate class	Fixed	All catchments within the same climatic group as the ungauged catchment are selected as donors. Donor selection is based on the catchment classification of Berghuijs et al. (2014) in which aridity, precipitation seasonality, and snowfall fraction are used for classification (see Figure A1 for a map of the classification). Ungauged catchments that did not belong to any class were excluded from the evaluation.
Random	Flexible	Three randomly selected catchments from all 670 potential donor catchments are used as donors.
Attr	Flexible	The three catchments most similar to the ungauged catchment in terms of attributes are selected as donors. Similarity was defined as the Eucledean distance calculated from area, aridity, precipitation seasonality, snowfall fraction, wetland fraction, clay fraction, and forest fraction. This approach is also known as the attribute-similarity approach or the physical-similarity approach (e.g., McIntyre et al., 2005; Oudin et al., 2010; Zhang & Chiew, 2009).
Methods that include distance information		
Geogr class	Fixed	All catchments within the same geographic area as the ungauged catchment are selected as donors. Donor selection is based on the major watershed regions (HUC2) defined by the U.S. Geological Service (USGS, 2020; see Figure A1 for a map of the classification).
Dist	Flexible	The three catchments that are spatially closest to the ungauged catchment are selected as donors. The coordinates of the catchment centroids were used to calculate the Euclidean distance between catchments. This approach is also known as the spatial-proximity approach (e.g., Lebecherel et al., 2016; Neri et al., 2020; Oudin et al., 2008; Parajka et al., 2005).
Dist & Attr	Flexible	The three catchments that are spatially closest (*Dist*) and most similar to the ungauged catchment in terms of attributes (*Attr*) are selected as donors.
Qsim transfer	Flexible	Simulated hydrographs are transferred from the three catchments that are spatially closest to the ungauged catchment.
Methods that include volume information		
WB class	Fixed	All catchments within the same water balance group as the ungauged catchment are selected as donors. Donor selection is based on a catchment classification, according to Schaller and Fan (2009) in which groups are defined based on the tendency to lose or gain groundwater (see Figure A1 for a map of the classification). Ungauged catchments that did not belong to any class were excluded from the evaluation.
Sign class	Fixed	All catchments within the same hydrological signature group as the ungauged catchment are selected as donors. Donor selection is based on an adapted version of the catchment classification of Jehn et al. (2020) in which runoff ratio, mean annual discharge, low flows (Q05), high flows (Q95), mean half-flow date, and recession slope are used for classification (see Figure A1 for a map of the classification). Ungauged catchments that did not belong to any class were excluded from the evaluation.
Regime class	Fixed	All catchments with the same hydrological regime as the ungauged catchment are selected as donors. Donor selection is based on the catchment classification of Brunner et al. (2020) in which the functional form of the hydrograph is used for classification (see Figure A1 for a map of the classification).

Table *2*
Continued

Approach		Source of parameter values
RMSE	Flexible	The three catchments with the smallest root mean square error for 12 observations in the (hypothetically) ungauged catchment are used as donors. The 12 observations were selected according to Pool et al. (2019) and included the annual peak and its five first recession days combined with observations at the 15th of every other month.
Vol	Flexible	The three catchments with the smallest volume error for the (hypothetically) ungauged catchment are used as donors.
Sign	Flexible	The three catchments that are most similar to the ungauged catchment in terms of hydrological signatures are selected as donors. Similarity was defined as the Eucledean distance calculated from runoff ratio, mean annual discharge, low flows (Q05), high flows (Q95), mean half-flow date, and recession slope.
Attr & Sign	Flexible	The three catchments that are most similar to the ungauged catchment in terms of attributes (*Attr*) and hydrological signatures (*Sign*) are selected as donors.
Methods that include distance and volume information		
Dist & Vol	Flexible	The three catchments that are spatially closest (*Dist*) and have the smallest volume error (*Vol*) for the ungauged basin are selected as donors.
Dist & RMSE	Flexible	The three catchments that are spatially closest (*Dist*) and are among the catchments with the smallest root mean square error for 12 observations (*RMSE*) in the ungauged catchment are selected as donors.
Dist & Sign	Flexible	The three catchments that are spatially closest (*Dist*) and most similar to the ungauged catchment in terms of hydrological signatures (*Sign*) are selected as donors.
Dist & Attr & Sign	Flexible	The three catchments that are spatially closest (*Dist*) and are most similar to the ungauged catchment in terms of attributes (*Attr*) and hydrological signatures (*Sign*) are selected as donors.
Methods without the transfer of parameter values		
Qobs transfer	Flexible	Observed hydrographs are transferred from the three catchments that are spatially closest to the ungauged catchment.

Note. The first column provides the label of each approach. The second and third columns provide detailed information about the selection of the donor catchments, that is, whether donors were selected from a predefined classification (fixed) or with a catchment-specific region of influence approach (flexible), as well as the physical attributes and hydrological signatures used for the selection. Approaches including volume information require either estimates of hydrological signatures (which can derived by regionalization) or a small number of local point observations.

catchment from October 1, 1989 to September 30, 1999. The individual simulations were then aggregated to an ensemble mean hydrograph by calculating their mean for each day. We used ensemble mean hydrographs rather than median hydrographs, because the ensemble mean conserves the total simulated discharge volume, and it was shown to outperform predictions with the single best parameter set (Neri et al., 2020; Seibert & Beven, 2009).

The comparison of all regionalization approaches was based on their simulation performance. Performance for the simulated ensemble mean hydrograph was assessed in terms of R_{KG}, R_{NP}, and each of their three error components (β, α_{KG}, and r_p for R_{KG} and β, α_{NP}, and r_S for R_{NP}). We also calculated the relative performance R^* (Seibert et al., 2018), which relates the performance of a particular regionalization approach (R_R) to the performance of an upper benchmark (R_U) and a lower benchmark (R_L). The relative performance was defined as $R^* = \left(R_R - R_L\right)/\left(R_U - R_L\right)$. The majority of the presented results are based on a model calibration with R_{NP}, and evaluate regionalization performance in terms of R_{NP} or R^*_{NP}, β, α_{NP}, and r_S. Calibration and evaluation results for R_{KG} are used in the robustness assessment. The absolute performance values for all calibrations and evaluations done in this study are provided in Data Set S2.

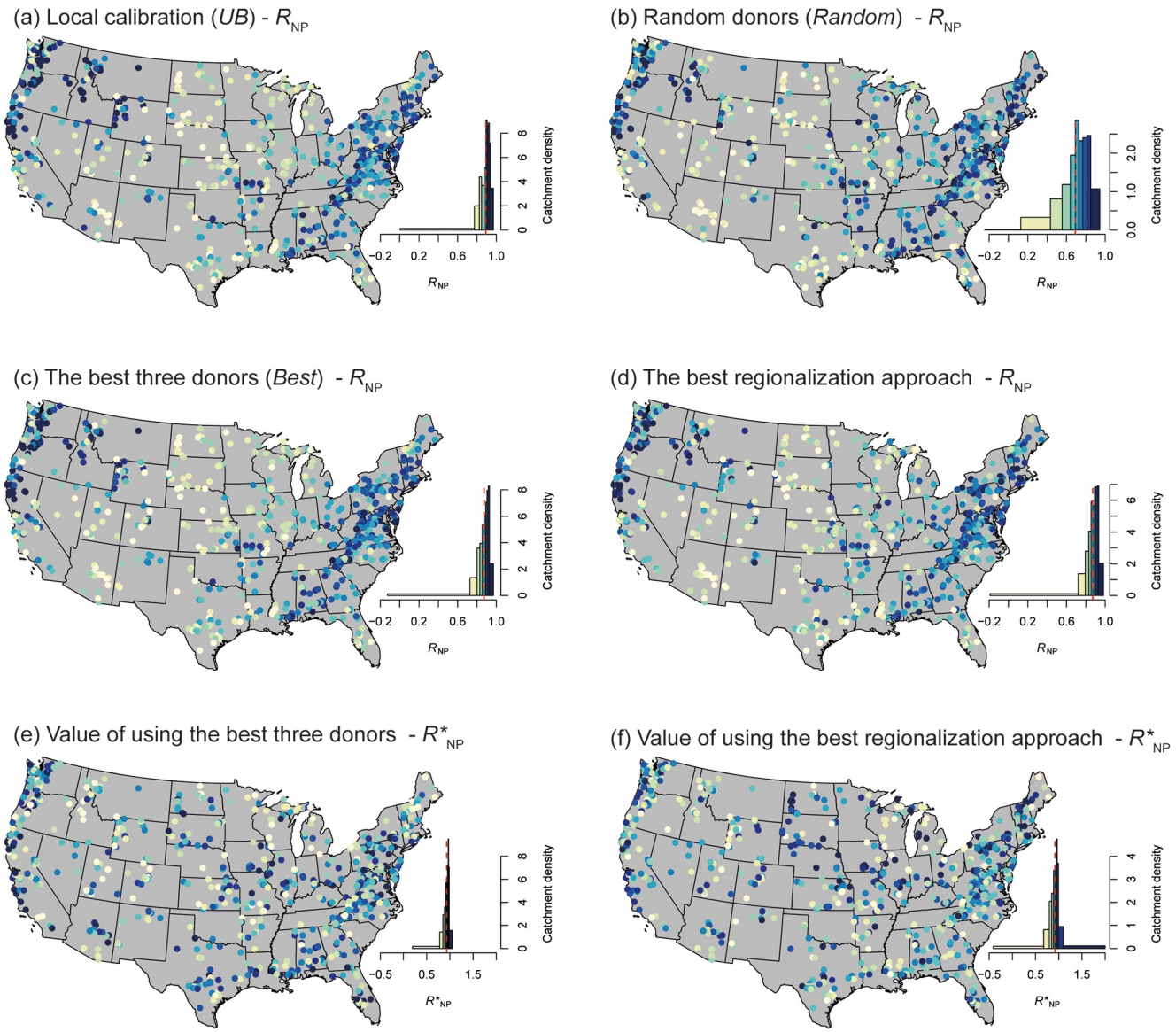

Figure 2. Spatial distribution of the model performance R_{NP} for (a) the local calibration (*UB*), (b) the regionalization with random donors (*Random*), (c) regionalization with the best three available donors (*Best*), (d) regionalization with the best of the tested regionalization approaches (note that the best approach is catchment-specific), (e) potential when using the best three available donors (*Best*), and (f) the potential when using the best of the tested regionalization approaches. Model performance values were grouped into 10 equally sized quantiles (that is, each group contains 1/10 of all 671 study catchments), which is indicated by the 10 colors. The red dashed line in the histogram indicates the median performance value. In (b), there were 44 catchments with R_{NP} between −0.2 and −11.6. In (e) and (f), R^*_{NP} corresponds to the percent improvement from a regionalization with random donors bounded by the local calibration.

2.3.3. Parameter Values and Sensitivity

In addition to the assessment of regionalization performance, we also investigated whether the parameter values and parameter sensitivity of donor catchments are linked to those of an ungauged receiver catchment. The parameter space analysis was conducted for each ungauged catchment and its best donor catchment (defined by R_{NP}). We first compared the parameter values of these two catchments by calculating the difference between the mean of the 10 calibrated parameter values. In a second step, we examined the parameter sensitivity of the donor and the receiver catchment. Parameter sensitivity was defined as the change in model performance (R_{NP}) when a given optimized parameter value was changed by ±2.5% while all other parameters were fixed. For each catchment, we then calculated the fractional sensitivity of

Table 3
Spearman Rank Correlation Between Model Performance and Catchment Characteristics

Physical attributes and hydrological signatures	R_{NP} *UB*	R_{NP} *Random*	R_{NP} *Best*	R_{NP} best approach	R^*_{NP} *Best*	R^*_{NP} best approach
Area	**−0.03**	**−0.10**	**−0.03**	−0.02	**0.05**	**0.08**
Aridity	**−0.53**	**−0.64**	**−0.61**	**−0.61**	−0.01	−0.02
Precipitation seasonality	**−0.41**	**−0.23**	**−0.37**	**−0.3**	**0.03**	**0.14**
Snowfall fraction	**0.21**	**0.08**	**0.12**	**0.14**	**−0.23**	−0.04
Wetland fraction	0.01	**0.03**	0	0	**−0.08**	0
Forest fraction	**0.47**	**0.47**	0.50	**0.52**	**−0.03**	0.02
Clay faction	**−0.18**	**−0.13**	**−0.13**	**−0.13**	**0.12**	0.03
Runoff ratio	**0.62**	**0.62**	**0.63**	**0.61**	**−0.15**	**−0.11**
Mean daily discharge	**0.63**	**0.67**	**0.67**	**0.65**	**−0.08**	−0.07
Low flows (Q5)	**0.46**	**0.49**	**0.44**	**0.43**	**−0.19**	**−0.12**
High flows (Q95)	**0.62**	**0.63**	**0.65**	**0.63**	**−0.08**	−0.07
Mean half-flow date	**−0.23**	**−0.3**	**−0.33**	**−0.32**	**−0.18**	−0.08
Recession slope	**−0.21**	−0.01	**−0.10**	**−0.05**	**0.22**	**0.17**

Note. Correlations were calculated for the local calibration (*UB*), the regionalization with random donors (*Random*), the regionalization with the best three available donors (*Best*), and the regionalization with the best of the tested regionalization approaches (note that the best approach is catchments-specific). Significant correlations ($p < 0.05$) are marked in bold letters.

a particular HBV model routine by taking the ratio of the sum of parameter sensitivities for a given routine and the sum of all parameter sensitivities.

3. Results and Discussion

3.1. Benchmarking Regionalization Performance in Different Geographic Regions

3.1.1. The Value of Using the Most Suited Regionalization Approach

With regionalization, one typically aims to reach high model performance values. Figures 2a–2d provide benchmark values for judging the goodness-of-fit of regionalization approaches within our study area. The best possible simulation performance was obtained by a local calibration (*UB*) with a median R_{NP} of 0.89 over all catchments. When using a simple form of regionalization, such as the use of random donor catchments (*Random*), median R_{NP} values were inferior (R_{NP} of 0.69) to this local calibration. However, regionalization performance could be relatively close to a local calibration when the best available donors were used for predicting streamflow in an ungauged catchment (*Best*). The median performance of these best donors (R_{NP} of 0.87) was identical to the performance of the best regionalization approach found for a given catchment. This similarity suggests that the best possible regionalization performance can be reached for many catchments with one of the regionalization approaches tested in this study.

The relative performance values R^*_{NP} presented in Figures 2e and 2f show the value of using the best possible donors or regionalization approach as opposed to using randomly selected donors. The best donors improved the performance for all catchments and for more than half of the catchments resulted in an improvement of at least 93% (Figure 2e). A comparable improvement (91%) was observed for the best regionalization approach found for a given catchment. However, its performance was worse than a regionalization with random donors in 8 out of the 671 catchments (Figure 2f). These results suggest that predictions in ungauged catchments can be surprisingly good if the most suitable donors are known. Similar conclusions were made by Beck et al. (2020) and Zhang and Chiew (2009), who compared an intelligent selection of donor catchments with a random selection of donors or parameter values. Since they were not explicitly searching for a regionalization benchmark but rather testing (new) approaches, their performance improvement (about 42%–68%) was smaller than the one reported here. Despite the general benefit of a good

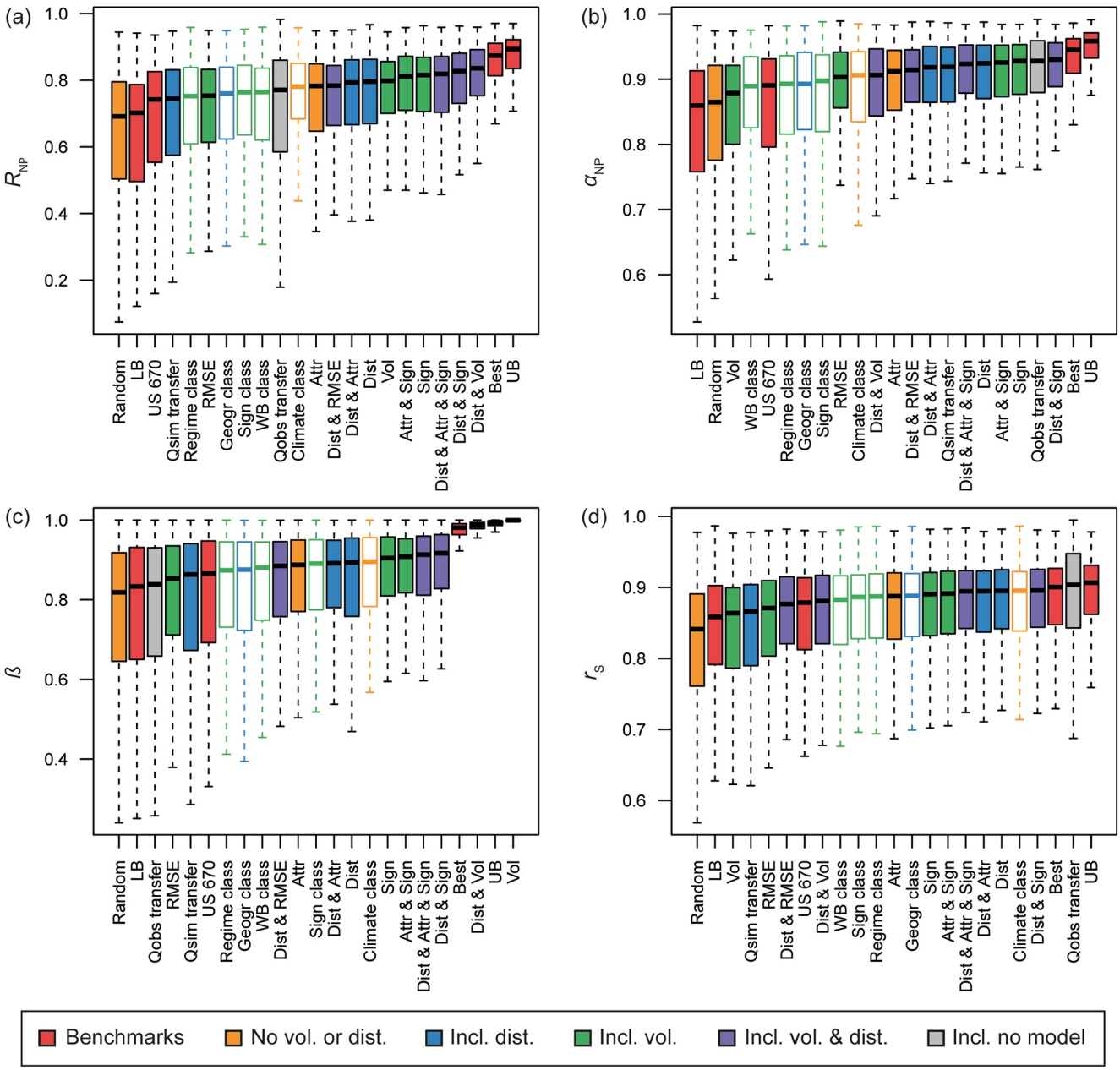

Figure 3. Model performance for the 671 catchments. Performance is shown for all tested regionalization approaches and benchmarks using (a) the total R_{NP} score as well as (b) its variability component α_{NP}, (c) its volume components β, and (d) its correlation component r_S. Note that the order of regionalization approaches on the x-axis is different in (a–d) as they are sorted by increasing median performance. Regionalization approaches are colored according to their category. Approaches based on catchment classification are highlighted by non-filled boxes.

regionalization approach, this study and Zhang and Chiew (2009) and Beck et al. (2020) found catchments for which random donors or parameter values could be the better choice. Reasons for this observation could be the coincidence of randomly selecting great donors (Zhang & Chiew, 2009) or the lack of attributes representing hydrological similarity (Oudin et al., 2010).

3.1.2. Spatial Patterns in Performance

The simulation performance R_{NP} strongly varied in space, and similar spatial performance patterns could be observed for the local calibration as for the regionalization approaches (Figures 2a–2d). To a large degree, the performance patterns followed the wetness gradients across the United States with the highest R_{NP} values in the Pacific Northwest, the northern Rocky Mountains, and the eastern United States (Table 3). Other catchment attributes such as area, snowfall, or soil properties showed no clear relationship with R_{NP} at the continental scale. The challenge of modeling relatively dry regions has been reported in several large-scale studies conducted in gauged (Arheimer et al., 2020; Mizukami et al., 2017; Poncelet et al., 2017) and hypothetically ungauged catchments (Beck et al., 2020; Parajka et al., 2013; Petheram et al., 2012) and was largely attributed to the higher flow variability and non-linearity of rainfall-runoff processes in arid catchments (Parajka et al., 2013; Petheram et al., 2012; Poncelet et al., 2017; Seibert et al., 2018). However, geological, topographic, or land-use characteristics can be important controls on regionalization performance within climatically similar regions of the United States (Singh et al., 2014) and may explain the small-scale performance differences observed in Figures 2a–2d.

While the regionalization performance R_{NP} was catchment-specific, there was no evidence for a spatial organization of the relative performance R^*_{NP} (Figures 2e and 2f). The relationships between R^*_{NP} and the catchment attributes or signatures were consistently weak (Table 3). This suggests that the search for the best possible regionalization approach is of comparable value for many catchments independent of how well they can be modeled. Similarly, the findings of Beck et al. (2020) indicate that the benefit of regionalization compared to using random parameters is not dependent on climate or topography. While their results were based on a distributed regression-based regionalization, our results represent the best possible regionalization of entire parameter sets.

3.2. In Search of the Best Regionalization Approach

3.2.1. A Ranking of Regionalization Approaches at the Continental Scale

We start the search for the best regionalization approach by comparing model performance at the continental scale (Figure 3). Simulations for all 671 catchments indicate that the performance difference between catchments is larger than between regionalization approaches. This is because most approaches have the potential to select donors that result in a model performance close to the upper benchmarks (*UB* and *Best*) for a limited number of catchments. However, the regionalization approaches differ considerably in their poorest performance values leading to clear differences in the suitability of the approaches for prediction in ungauged basins.

Figure 3a presents the ranking of regionalization approaches in simulating a range of hydrograph aspects (R_{NP}; an equivalent figure for R_{KG} is provided in Figure S4). Generally, the most promising and most widely applicable regionalization approaches combine information on spatial distance and volume for selecting donor catchments (median R_{NP} of 0.82 and median R_{KG} of 0.67). Approaches based on a single type of information (i.e., attributes, distance, or volume) and a flexible region of influence were located in the middle ranks. Volume seemed to be the best similarity indicator (median R_{NP} of 0.80 and median R_{KG} of 0.66), followed by spatial proximity (median R_{NP} of 0.79 and median R_{KG} of 0.64) and catchment attributes (median R_{NP} of 0.78 and median R_{KG} of 0.62). Using the same information to select donors from a fixed catchment classification instead of a catchment-specific similarity region typically resulted in lower performances (median R_{NP} of 0.76 and median R_{KG} of 0.58). Yet, classification-based regionalization approaches were still considerably better than the lower benchmarks or a random selection of donors (median R_{NP} of 0.7 and median R_{KG} of 0.49).

There are some remarkable changes in the ranking described above if one looks at individual hydrograph aspects (Figures 3b–3d). For example, when simulating flow variability (α_{NP}) volume-based approaches were most successful, which is reflected in the high rank of approaches considering hydrological signatures (*Sign* or *Attr & Sign*) or the high rank of *Qobs transfer*. The ranking for flow volume (β) was similar to that of the entire hydrograph, with the main difference being that climatic aspects (*Climate class*) could become a surprisingly good predictor for catchment similarity. Finally, using spatially close donors (*Qobs transfer*,

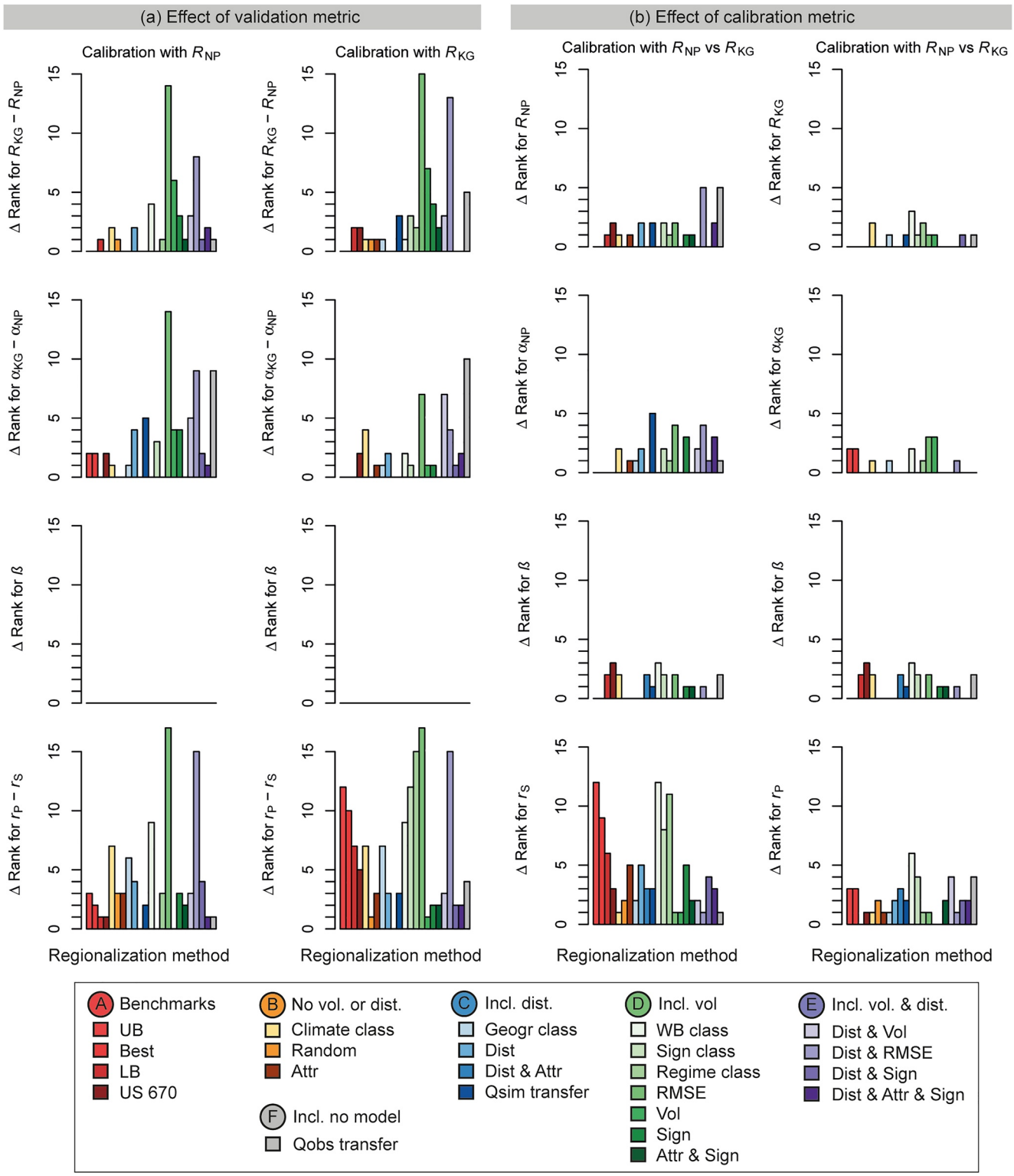

Figure 4. Robustness of the ranking of regionalization approaches. (a) Effect of a validation metric: robustness is defined as the change in rank if a different evaluation metric is used for a given calibration metric. The values for β are zero as β is identical for R_{NP} and R_{KG}. (b) Effect of a calibration metric: robustness is defined as the change in rank for a given evaluation metric if a different calibration metric is used. Note that the maximum possible change in rank is 21.

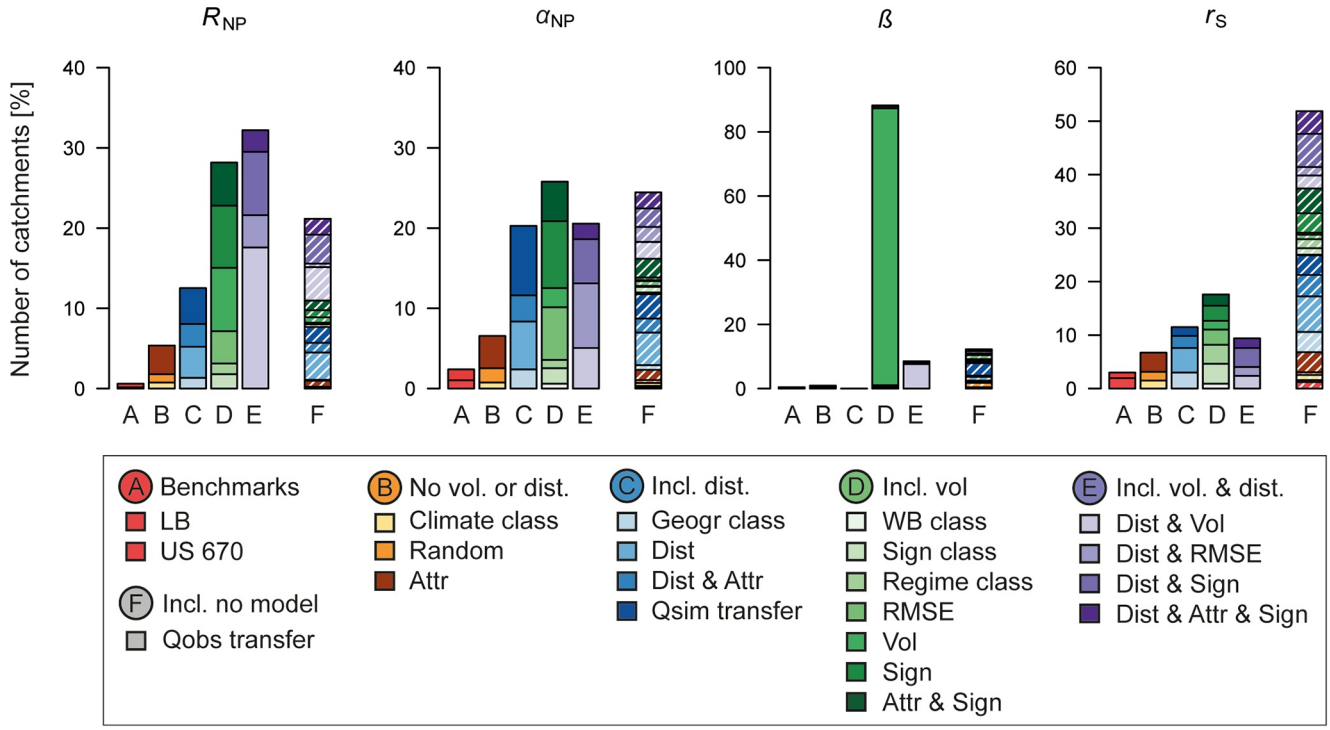

Figure 5. The best regionalization approach for each of the 671 catchments as a function of the performance metric (R_{NP}, α_{NP}, β, and r_s). Regionalization approaches are grouped by category. In the case that *Qobs transfer* was the best regionalization approach (F), the colors of the bar indicate the second best approach. Note the different y-axis scales.

all *Dist* approaches, or even *Climate class*) tended to be more important than using a volume-based donor selection for simulating streamflow dynamics (r_S).

The general tendency of the suitability of regionalization approaches found at the continental scale agrees with previously reported results. Similar to the studies of Merz and Blöschl (2004), Oudin et al. (2008), Swain and Patra (2017), or Zhang and Chiew (2009), we found that distance-based approaches tend to outperform attribute-based approaches. This could be linked to the relatively high gauging station density in many humid regions of the United States (Lebecherel et al., 2016; Neri et al., 2020) or the lack of catchment attributes representing a wide range of catchment responses over large scales (Merz et al., 2020; Oudin et al., 2010; Singh et al., 2014). In fact, Jehn et al. (2020) found that catchment attributes (excluding climatic aspects) can vary considerably within hydrologically similar regions in the United States and can often not sufficiently explain hydrological behavior. In contrast to catchment attributes or geographical information, basic hydrological information can provide a more direct measure of catchment similarity. While the high performance of volume-based regionalization approaches is not surprising, the implementation of such approaches can be restricted by data availability. However, it has been shown that even a small number of point observations (Pool et al., 2019; Rojas-Serna et al., 2016; Viviroli & Seibert, 2015) or regionalized signatures (Masih et al., 2010) can significantly improve predictions, which makes volume-based approaches an interesting alternative to commonly used approaches. Our results further suggest that donors selected from a flexible region of influence are better predictors for the ungauged catchment than donors from a prior classification. This can be explained partly by the deterministic nature of the classifications (Burn, 1990) and a large number of (donor) catchments within each class. With an increasing number of catchments, the dissimilarity among catchments likely increases leading to decreased performance (Arsenault & Brissette, 2014; Neri et al., 2020; Oudin et al., 2008; Yang et al., 2018). This is in agreement with Figure 1, which shows that the number of nearly perfect donors is limited, causing a large class to include, per definition, several less suitable donors.

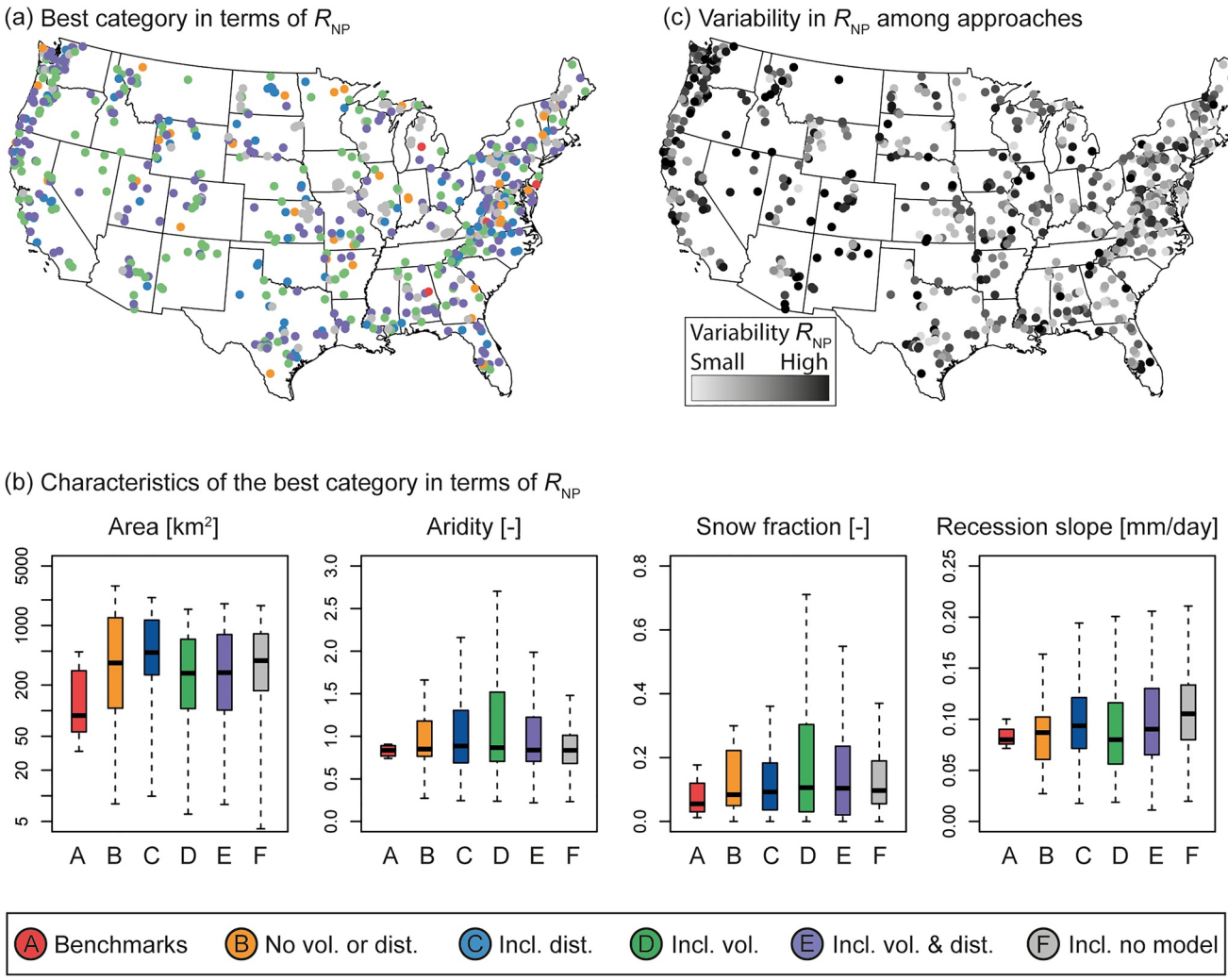

Figure 6. (a) Spatial distribution and (b) selected catchment characteristics of the best regionalization category in terms of R_{NP}, (c) Spatial distribution of the variability (standard deviation) in R_{NP} among all tested regionalization approaches.

In contrast to the abundant number of studies comparing different regionalization approaches, studies that can be used to compare the absolute model performance in terms of R_{KG} (rather than in terms of the Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) or adapted versions of R_{KG} as used by Beck et al., 2020 and Merz et al., 2020) are still limited. Among these studies, Neri et al. (2020) found median R_{KG} values between 0.6 and 0.75 for 209 Austrian catchments regionalized with attribute similarity or spatial proximity, which are similar performance values to the ones of this studies. Our performance values are, however, high compared to the global studies of Arheimer et al. (2020) and Beck et al. (2016), who used regressions and process-dependent classifications for regionalizing model parameter values and reported median monthly R_{KG} values of 0.32 and 0.45.

3.2.2. Robustness of the Ranking of Regionalization Approaches at the Continental Scale

Many regionalization studies compare several regionalization approaches with the aim to determine the most suitable approach for a given set of catchments. Figure 4 shows how much such a comparison (i.e., ranking) of regionalization approaches could be influenced by the choice of performance metrics used for calibrating the hydrological model and evaluating its simulations. Our results suggest that the choice of the evaluation metric (Figure 4a) tends to have a stronger impact on the ranking than the use of different calibration metrics (Figure 4b). More specifically, calibrating against R_{NP} or R_{KG} changed the performance of a

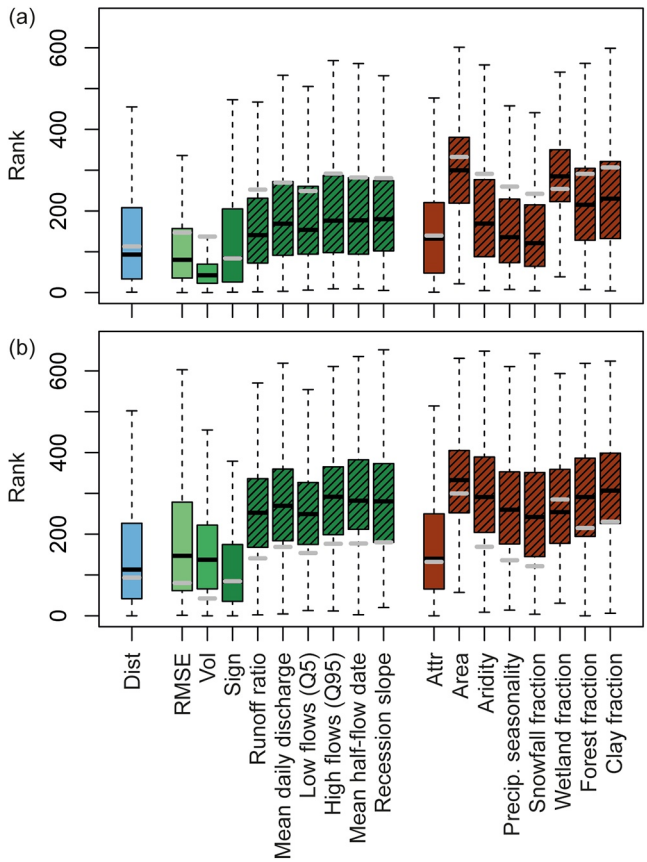

Figure 7. (a) Similarity of the three best available donors and the ungauged catchments. Similarity was defined as the average rank of the best three donors among all 671 catchments in the geographical, signature, and attribute space. (b) Regionalization performance of donors selected using similarity in spatial distance, signatures, or attributes. Performance (R_{NP}) was evaluated by the average rank of the most similar donors among all 671 catchments. The gray line on top of each boxplot in (a) indicates the median value of the corresponding boxplot in (b) and vice versa.

particular regionalization approach by about 1.5 ranks. In contrast, evaluating regionalization performance with R_{NP} (including α_{NP} and r_S) or R_{KG} (including α_{KG} and r_P) resulted in an average change of 3.2 ranks. Furthermore, the results indicate that the ranking of regionalization approaches is likely to be more robust when considering multiple hydrograph aspects (R_{NP} or R_{KG}) instead of a single hydrograph aspect (β, α_{NP}, α_{KG}, r_S, or r_P). The ranking based on streamflow dynamics (r_S, or r_P) was particularly sensitive to the choice of the performance metric. The described effect of performance metrics on the ranking of regionalization approaches varied among approaches. There was a slight tendency toward a more consistent ranking if distance was included in donor selection (e.g., *Dist & Attr*, *Dist & Sign*, *Dist & Attr & Sign*). The least robust ranking was observed for volume- or classification-based approaches (e.g., *RMSE or RMSE & Dist*, *WB Class or Sign Class*) and the direct transfer of observed hydrographs from nearby stations (*Qobs transfer*).

A major outcome of the robustness assessment is that the suitability of a regionalization approach is dependent on the hydrograph aspects that we aim to simulate. Similar findings were reported by Singh et al. (2014), who showed that the choice of a performance metric could change the importance of catchment attributes as a proxy for hydrological similarity. An evaluation of multiple hydrograph aspects is therefore crucial when ranking regionalization approaches and will likely facilitate the comparability of results from different regionalization studies. To address the uncertainty related to the choice of regionalization approaches, it may be worth exploring the benefits of an ensemble regionalization approach. For example, Farmer and Vogel (2013) and Waseem et al. (2015) showed that performance-weighted ensembles often outperform individual approaches to predict streamflow signatures in ungauged basins. More recently, Razavi and Coulibaly (2016) and Swain and Patra (2019) successfully implemented an ensemble regionalization approach to predict continuous streamflow in ungauged basins. While these studies typically applied a time-invariant weighting approach, Razavi and Coulibaly (2016) chose to weight the simulations based on the daily performance in the gauged catchment. Our results suggest that a dynamic ensemble weighting may also be applied in prediction mode if regionalization approaches were weighted based on their performance for a particular flow condition.

3.2.3. The Best Regionalization Approach From a Catchment Perspective

The search for the best regionalization approach from a catchment perspective (that is, in a bottom-up approach) allows refining the performance trends observed at the continental scale. While the more local perspective (Figure 5; see Figure S5 for example hydrographs) supports the continental-scale tendencies (Figure 3), it also indicated that methods considered less suitable, if evaluated for all 671 catchments, could be the best choice for an individual catchment. Examples that support the general trends are the regionalization approaches *Dist & Vol*, *Dist & Sign*, and *Vol* or *Sign*. These regionalization approaches were the best approaches transferring entire parameters sets in terms of R_{NP} for 22%, 12%, and 9% of the catchments, respectively. In contrast, results for the approaches based on a signature classification (*Sign class*), a few point observations (*RMSE*), or attributes (*Attr*) indicate that the search for a best regionalization approach is to some degree catchment dependent. Using these approaches to select donors could be the best option in terms of R_{NP} for 2%–4% of the ungauged catchments despite their relatively poor performance at the continental scale. A further remarkable observation was that the approach based on the direct transfer of observed streamflow from nearby catchments, *Qobs transfer*, had a varying performance for R_{NP}, α_{NP}, β, and r_S at continental scale but was the most competitive option at the catchment scale as it was the best approach for 12%–52% of the catchments.

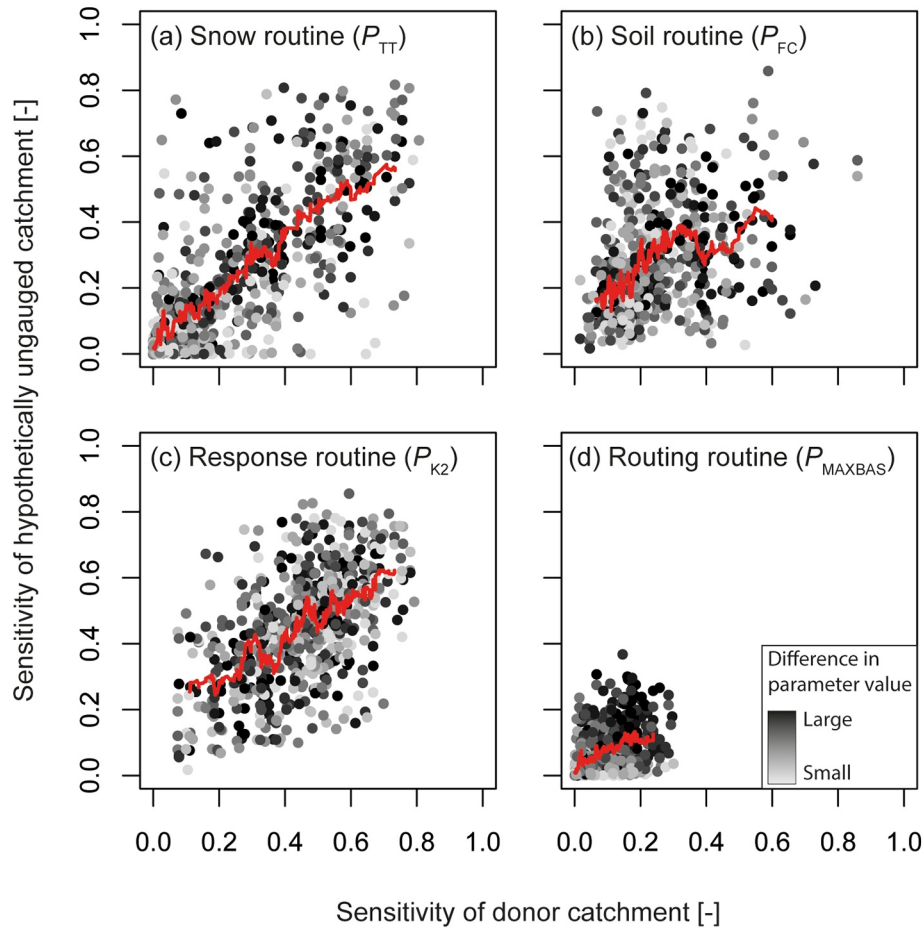

Figure 8. Parameter sensitivity and parameter values of the 671 hypothetically ungauged catchments and their best donor catchment. The subplots show the fraction of total parameter sensitivity contributed by parameters of (a) the snow routine, (b) the soil routine, (c) the response routine, and (d) the routing routine. The gradation (white to black) of the points indicates the difference in mean parameter values of the hypothetically ungauged basin and its best donor catchment. These differences are shown for the most sensitive parameter of each model routine. P_{TT}, P_{FC}, P_{K2}, and P_{MAXBAS} are the threshold temperature for snowfall, the maximum soil moisture storage, the storage (or recession) coefficient for interflow, and the length of the triangular weighting function. The red line is the average over a moving window of 29 values.

Given that the best regionalization approach varies among the 671 catchments it is essential to understand where a particular approach performs best. However, there was no evidence for a spatial clustering of the best regionalization approaches (or category of approaches; Figure 6a) or a strong relationship with catchment characteristics (Figure 6b). Similar to the findings of He et al. (2011) and Razavi and Coulibaly (2013), our results indicate that predicting and choosing the best regionalization approach for an ungauged catchment is still a major challenge. The transfer and interpolation of observed daily streamflow from nearby catchments (*Qobs transfer*) could therefore provide an interesting and valuable alternative to the transfer of model parameters if the causal relationship between precipitation and streamflow is not of interest (Parajka et al., 2015; Patil & Stieglitz, 2012; Razavi & Coulibaly, 2016). Furthermore, using the best possible regionalization approach is most important if performance differences among approaches are expected to be relatively large. Figure 6c shows that these differences varied considerably between catchments. Regionalization performance thereby tended to be more dissimilar in the poorer modeled catchments (Figures 2c and 2d), which confirms results previously reported by Zhang and Chiew (2009). Consequently, choosing a good regionalization approach (or avoiding the use of an unsuitable one) is most challenging for poorly modeled catchments — however, the benefit of knowing the best possible regionalization approach is of comparable value for many catchments, independent of how well they can be modeled (Figures 2d–2f).

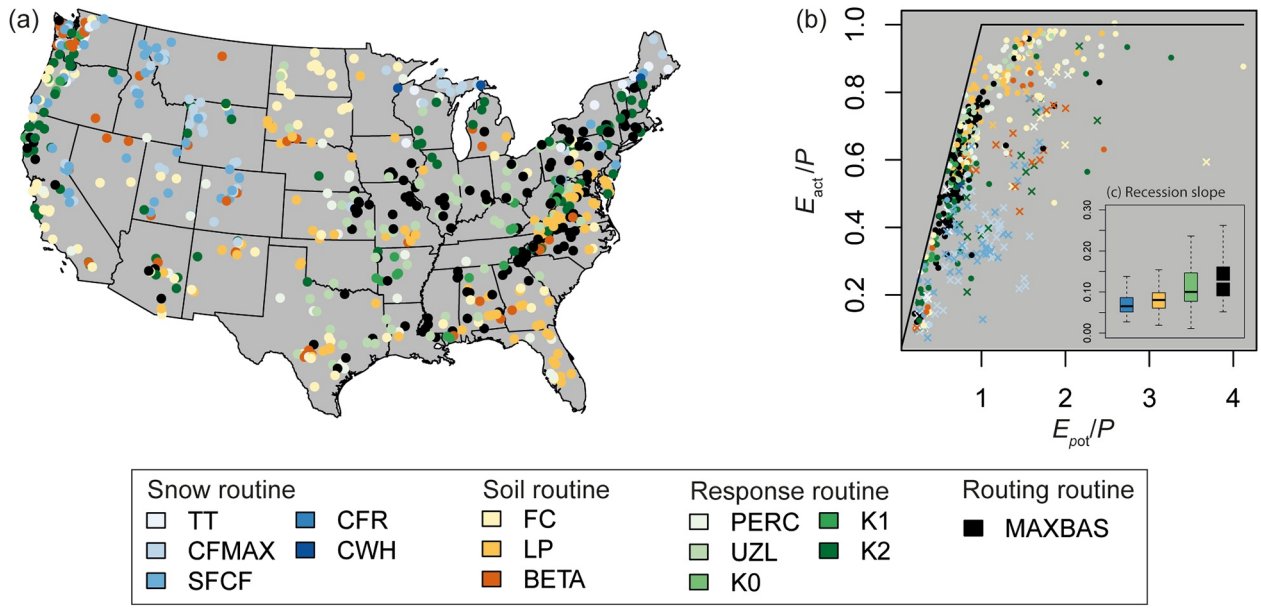

Figure 9. (a) Spatial distribution of the most sensitive parameter within the most sensitive HBV model routine. (b) The most sensitive parameter of the most sensitive routine plotted in the Budyko-space (E_{act} is actual evapotranspiration, E_{pot} is potential evapotranspiration, and P is precipitation). Catchments with more than 25% of precipitation falling as snow are marked with a cross. (c) Recession slope of catchments for which the snow routine (blue), soil routine (yellow), response routine (green), or routing routine (black) was the most sensitive HBV model routine. Note that the most sensitive HBV model routine was defined by taking the total sensitivity of a routine normalized by the number of parameters in that routine.

3.3. Learning From the Best Donor Catchments

3.3.1. Importance of Catchment Characteristics for Selecting the Best Donor Catchments

We hypothesized that the catchment characteristics of the three best donor catchments available in the data set (*Best*) could help to identify important aspects for selecting suitable donors. To address this hypothesis, we first analyzed the similarity of each ungauged catchment and its three best donors in the geographical, signature, and attribute space using the Euclidean distance. The similarity for different characteristics was compared by ranking the similarity values of all potential donors and calculating the average rank of the three best donors (Figure 7a). Catchment similarity was generally highest for the hydrological signatures, followed by the spatial distance, and was smallest for the physical attributes. Similarity was typically higher for a combination of several signatures (*Sign*) and attributes (*Attr*) than for individual aspects (e.g., runoff ratio or area). Among the hydrological aspects, similarity was particularly high for streamflow volume (*Vol*), indicating that the three best donors resulted in a low relative error in streamflow predictions in the ungauged catchment. In a second step, we used spatial distance, signatures, or attributes to select donors and evaluated their performance in terms of R_{NP}. Similar as for the characteristics, we calculated the average performance rank of the three most similar donor catchments (Figure 7b). The ranks of these most similar donors in the "performance space" are typically clearly worse than the ranks of the three best donors in the "characteristics space." The smallest rank differences were observed for the combinations of *Sign* or *Attr*, and *Dist*.

Our findings indicate that using (seemingly) essential catchment characteristics for selecting donor catchments does not necessarily lead to a good regionalization performance. One reason for this finding could be the spatial variability of the importance of catchment characteristics. For example, within the United States, geology and topography are important characteristics in humid mountainous regions, whereas land use is the most important proxy for hydrological similarity in humid plains (Singh et al., 2014). Furthermore, the study of Bárdossy et al. (2016) conducted with catchments in the eastern United States suggests that the temporal variability in catchment similarity could pose a considerable challenge for selecting a universal set of donor catchments. Given this variability, an improved understanding of catchment similarity at multiple temporal and spatial scales will be a foundation for advancing regionalization (He et al., 2011; Wagener

et al., 2007). The search for new hydrologically relevant characteristics that describe regional and local hydrological processes (Merz et al., 2020) occurring at the surface and the subsurface (Oudin et al., 2010) could play a key role.

A potentially interesting point to explore more in a future study is whether catchment characteristics could be used to exclude donors from the pool of potential donors, that is, switching from selecting suitable donors to a (combined) strategy where catchments are deselected based on their characteristics. Although streamflow volume, for example, seems to be an important characteristic of well-performing donor catchments, using it as a selection criterion is no guarantee for the suitability of the selected donors. However, since the best performing donor catchments rank high with respect to streamflow volume, it seems that poor streamflow volume performance is at least a guarantee for not being a suitable donor.

3.3.2. Parameter Values and Sensitivity of the Best Donor Catchments

A successful regionalization essentially relies on finding donor catchments that share similar parameter values as the ungauged target catchment. We, therefore, investigated to which extent the parameter values and the parameter sensitivity of the single best donor catchment are related to those of the ungauged receiver catchment (Figure 8). Despite the considerable scatter in parameter sensitivity, the hypothetically ungauged catchments and their best donor catchment tended to have a similar sensitivity for a particular HBV model routine. The relationship was strongest for the snow routine (Spearman rank correlation coefficient ρ equals 0.75), followed by the response and routing routine (ρ equals 0.55 and 0.53, respectively), and was weakest for the soil routine (ρ equals 0.48). While a similar sensitivity seemed to be important for a successful parameter transfer, the values of the most sensitive parameter of a particular routine could be very different. The reason for this could be the dependency of parameter values (Arsenault & Brissette, 2014; Bárdossy, 2007; McIntyre et al., 2005), which can be expected to be particularly pronounced for parameters within the same model routine.

Figure 9 shows the spatial distribution of the most sensitive model routine. As could be expected, the snow routine was the most sensitive routine in the Rocky Mountains, Sierra Nevada, or the Great Lakes Region where snow processes dominate the runoff regime. The soil routine was particularly important in water-limited regions such as the Great Plains and the Southwest, or in regions with high actual evaporation, such as the Florida Peninsula. Finally, the response routine and the routing routine were typically the most sensitive routines in energy-limited regions without considerable snowfall. This includes large regions of the Eastern United States and the Pacific Northwest. Within these regions, simulations tended to be more sensitive to the routing routine than the response routine if the recession slope was rather steep. The spatial distribution of the most sensitive HBV model routine across the United States confirms the findings of several HBV parameter sensitivity studies conducted at a local scale (Abebe et al., 2010; Medina & Muñoz, 2020; Pianosi & Wagener, 2016; Zelelew & Alfredsen, 2013) and reflects the large-scale variability in dominant hydrological processes. Applying the spatially distributed PRMS model in the United States, Markstrom et al. (2016) used parameter sensitivity analysis to derive detailed maps of dominant hydrological processes. While such maps are likely affected by the spatial resolution of a model, the temporal working scale, or the performance metric used to quantify sensitivity, they could be used for catchment classification (Markstrom et al., 2016). Indeed, the spatial pattern of the HBV model routine importance observed in this study followed the streamflow-based regime classification from Brunner et al. (2020). Their melt regime with spring and summer snowmelt floods is in the region of high snow routine importance, their intermittent regime with arid catchments and irregular precipitation events coincides with the region of high soil routine importance, and their precipitation-dominated regimes (winter regimes and New-Year's regime) are regions of high response routine or routing routine importance. While parameter-similarity has previously been proven to result in high regionalization performance (Parajka et al., 2005), our findings further indicate that it would be interesting to explore the value of parameter-sensitivity classifications for the regionalization of model parameters to ungauged basins.

4. Conclusions

A systematic assessment of 19 regionalization approaches was conducted using 671 U.S. catchments with a large variation in hydroclimatology and a homogenized modeling protocol (i.e., identical model structure, as well as calibration and evaluation processes). The assessment includes the introduction of a regionalization benchmark to evaluate model performance across hydroclimates, the testing of the robustness of the ranking of regionalization approaches under different evaluation metrics, and the description of the best possible donor catchments in the attribute and parameter space. The main findings of this study can be summarized in three lessons:

Lesson 1 — Catchments in any geographic region can benefit from a well-chosen regionalization approach: The performance of streamflow simulations with both locally calibrated parameter sets and regionalized parameter sets strongly varied in space, whereby higher performance values could be achieved with increasing wetness of a catchment. However, the relative performance (defined as the performance relative to an upper (local calibration) and lower (randomly chosen donors) benchmark) of the best regionalization approach for each catchment was generally high without a clear spatial pattern. In other words, catchments with very different hydrological characteristics can equally benefit from the careful choice of a regionalization approach.

Lesson 2 — Almost perfect donors exist and an excellent relative model performance can be reached for most catchments with current regionalization approaches: Besides using an upper benchmark as a measure of how well a catchment can perform in general, we propose the use of a regionalization benchmark as a reference of what can be achieved at best when using regionalized parameter sets. Here, we used the best three donor catchments as a regionalization benchmark. Predictions based on these best three donors reached a performance close to a local calibration, indicating that almost perfect donors exist for most catchments. Also, the best regionalization approach per catchment performed almost as well as the best three donors, suggesting that regionalization strategies can identify suitable donors. Quantifying the characteristics of excellent donor catchments and defining a universal regionalization strategy to identify these donor catchments remains challenging. Yet, our analyses highlighted that there is considerable potential for improvement in the prediction in ungauged catchments.

Lesson 3 — The ranking of regionalization approaches depends more on how the predicted hydrographs are evaluated than on how the donor catchments are calibrated: Our findings revealed that the ranking of regionalization approaches, and the search for the best donor catchments, depends on which hydrograph aspects are considered. The choice of the evaluation metric tended to be of higher importance than the choice of the calibration metric. We, therefore, recommend making a multi-criteria evaluation an integral part of any comparative assessment of regionalization performance.

From a scientific perspective, our study showed that an excellent (relative) performance can be reached for most catchments with currently used regionalization approaches. The major challenge remains choosing the most suitable approach, whereby an improved understanding of the characteristics and parameter values of great donors and their relationship to an ungauged catchment will be key to advance regionalization further. From a practical perspective, our study suggests that distance and some volume information are among the most promising selection criteria for donor catchments. While volume information is per definition missing in ungauged catchments, it could be derived from estimated signatures or a small number of field measurements. If predictions under relatively stationary conditions are of interest, the transfer of observed daily streamflow from nearby catchments provides an interesting and valuable alternative to the transfer of model parameters.

Appendix A: Catchment Classifications and Geographic Regions

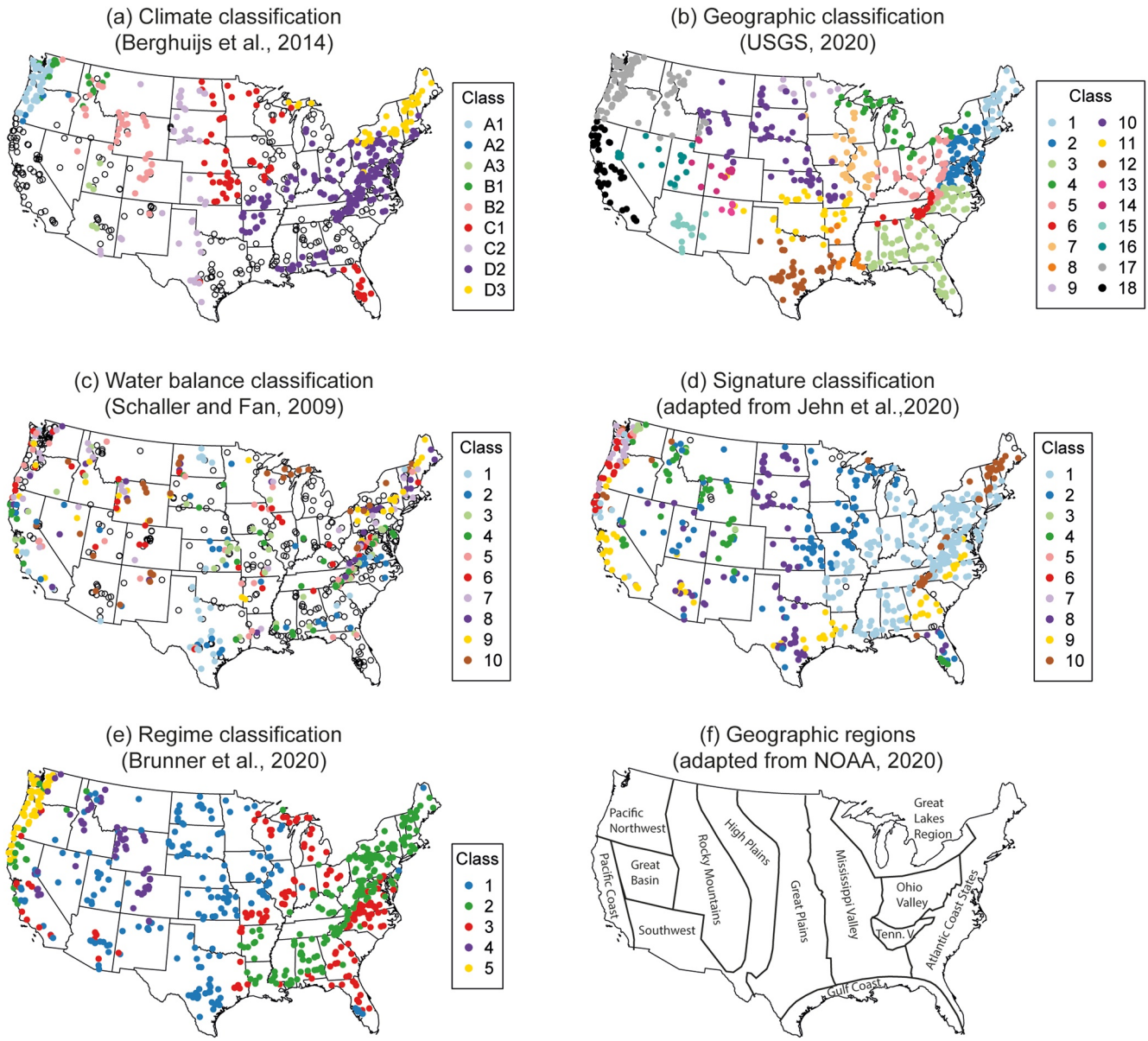

Figure A1. (a–e) Catchment classifications used for regionalization in this study. The spatial distribution of classes is shown in colors. Empty circles indicate catchments that could not be classified because they were not included in the reported classification (c and d), or because they could not be assigned to any class using the reported class boundaries (a). The number of unclassified catchments is 207, 0, 332, 28, and 0 for the climate, geographic, water balance, signature, and regime classification, respectively. (f) Geographic regions of the United States (NOAA, 2020).

Data Availability Statement

The data used in this study are publicly available. The hydrometeorological time series and most catchment characteristics were retrieved from Newman et al. (2015) and Addor et al. (2017). SRTM elevation data were extracted from Jarvis et al. (2008) and the spatial extent of wetlands was extracted from Lehner and Döll (2004). The streamflow recession slope was calculated with the *EflowStats* R-package (USGS, 2014). Data on wetland fraction and recession slope for each catchment are provided in the Data Set S1. The model performance value for each catchment, regionalization approach, calibration, and evaluation criteria are provided in the Data Set S2.

Acknowledgments
The authors thank Florian Jehn for the valuable discussions in the early phase of this project and for sharing his expertise on classification. Open access funding provided by ETH-Bereich Forschungsanstalten.

References

Abebe, N. A., Ogden, F. L., & Pradhan, N. R. (2010). Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: Implications for parameter estimation. *Journal of Hydrology*, *389*(3–4), 301–310. https://doi.org/10.1016/j.jhydrol.2010.06.007

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J., Hasan, A., & Pineda, L. (2020). 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

Arsenault, R., & Brissette, F. P. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, *50*(7), 6135–6153. https://doi.org/10.1002/2013wr014898

Bao, Z., Zhang, J., Liu, J., Fu, G., Wang, G., He, R., et al. (2012). Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions. *Journal of Hydrology*, *466*, 37–46. https://doi.org/10.1016/j.jhydrol.2012.07.048

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Bárdossy, A., Huang, Y., & Wagener, T. (2016). Simultaneous calibration of hydrological models in geographical space. *Hydrology and Earth System Sciences*, *20*(7), 2913–2928. https://doi.org/10.5194/hess-20-2913-2016

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17), e2019JD031485. https://doi.org/10.1029/2019jd031485

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599–3622. https://doi.org/10.1002/2015wr018247

Berghuijs, W. R., Sivapalan, M., Woods, R. A., & Savenije, H. H. (2014). Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, *50*(7), 5638–5661. https://doi.org/10.1002/2014wr015692

Bergström, S. (1976). *Development and application of a conceptual runoff model for scandinavian catchments* (pp. 134). Sweden: SMHI, Norrköping.

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, *249*(1–4), 11–29. https://doi.org/10.1016/s0022-1694(01)00421-8

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *19*(11), 4559–4579. https://doi.org/10.1002/hyp.3360090305

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the United States: Assessment using functional classification. *Hydrology and Earth System Sciences*, *24*(8), 3951–3966. https://doi.org/10.5194/hess-24-3951-2020

Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, *26*(10), 2257–2265. https://doi.org/10.1029/wr026i010p02257

de Lavenne, A., Andréassian, V., Thirel, G., Ramos, M. H., & Perrin, C. (2019). A regularization approach to improve the sequential calibration of a semidistributed hydrological model. *Water Resources Research*, *55*(11), 8821–8839. https://doi.org/10.1029/2018wr024266

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Farmer, W. H., & Vogel, R. M. (2013). Performance-weighted methods for estimating monthly streamflow at ungauged sites. *Journal of Hydrology*, *477*, 240–250. https://doi.org/10.1016/j.jhydrol.2012.11.032

Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R., & Tollan, A. (1979). Hydrologic regions in the Nordic countries. *Hydrology Research*, *10*(5), 273–286. https://doi.org/10.2166/nh.1979.0010

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

He, Y., Bárdossy, A., & Zehe, E. (2011). 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

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Jarvis, A., Reuter, H., Nelson, A., & Guevara, E. (2008). *Hole-filled SRTM for the globe Version 4*. Retrieved from http://srtm.csi.cgiar.org

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081–1100. https://doi.org/10.5194/hess-24-1081-2020

Johansson, B. (2000). Areal precipitation and temperature in the Swedish mountains: An evaluation from a hydrological perspective. *Hydrology Research*, *31*(3), 207–228. https://doi.org/10.2166/nh.2000.0013

Kokkonen, T. S., Jakeman, A. J., Young, P. C., & Koivusalo, H. J. (2003). Predicting daily flows in ungauged catchments: Model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. *Hydrological Processes*, *17*, 2219–2238. https://doi.org/10.1002/hyp.1329

Lebecherel, L., Andréassian, V., & Perrin, C. (2016). 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

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998wr900018

Lehner, B., & Döll, P. (2004). 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

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). 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

Madsen, H. (2000). Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology*, *235*(3–4), 276–288. https://doi.org/10.1016/s0022-1694(00)00279-1

Markstrom, S. L., Hay, L. E., & Clark, M. P. (2016). Towards simplification of hydrologic modeling: Identification of dominant processes. *Hydrology and Earth System Sciences*, *20*(11), 4655–4671. https://doi.org/10.5194/hess-20-4655-2016

Masih, I., Uhlenbrook, S., Maskey, S., & Ahmad, M. D. (2010). Regionalization of a conceptual rainfall–runoff model based on similarity of the flow duration curve: A case study from the semi-arid Karkheh basin, Iran. *Journal of Hydrology*, *391*(1–2), 188–201. https://doi.org/10.1016/j.jhydrol.2010.07.018

McIntyre, N., Lee, H., Wheater, H., Young, A., & Wagener, T. (2005). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, *41*(12), W12434. https://doi.org/10.1029/2005wr004289

Medina, Y., & Muñoz, E. (2020). Analysis of the relative importance of model parameters in watersheds with different hydrological regimes. *Water*, *12*(9), 2376. https://doi.org/10.3390/w12092376

Merz, R., & Blöschl, G. (2004). Regionalisation of catchment model parameters. *Journal of Hydrology*, *287*(1–4), 95–123. https://doi.org/10.1016/j.jhydrol.2003.09.028

Merz, R., Tarasova, L., & Basso, S. (2020). Parameter's controls of distributed catchment models—how much information is in conventional catchment descriptors? *Water Resources Research*, *56*(2), e2019WR026008. https://doi.org/10.1029/2019wr026008

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. https://doi.org/10.1007/bf01897163

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017wr020401

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I - A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Neri, M., Parajka, J., & Toth, E. (2020). Importance of the informative content in the study area when regionalising rainfall-runoff model parameters: The role of nested catchments and gauging station density. *Hydrology and Earth System Sciences*, *24*(11), 5149–5171. https://doi.org/10.5194/hess-24-5149-2020

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

NOAA. (2020). Geographical reference maps. Retrieved from https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/geography

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). 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

Oudin, L., Kay, A., Andréassian, V., & Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, *46*(11), W11558. https://doi.org/10.1029/2009wr008887

Parajka, J., Merz, R., & Blöschl, G. (2005). A comparison of regionalisation methods for catchment model parameters. *Hydrology and Earth System Sciences*, *9*(3), 157–171. https://doi.org/10.5194/hess-9-157-2005

Parajka, J., Merz, R., Skøien, J. O., & Viglione, A. (2015). The role of station density for predicting daily runoff by top-kriging interpolation in Austria. *Journal of Hydrology and Hydromechanics*, *63*(3), 228–234. https://doi.org/10.1515/johh-2015-0024

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–Part 1: Runoff-hydrograph studies. *Hydrology and Earth System Sciences*, *17*(5), 1783–1795. https://doi.org/10.5194/hess-17-1783-2013

Patil, S., & Stieglitz, M. (2012). Controls on hydrologic similarity: Role of nearby gauged catchments for prediction at an ungauged catchment. *Hydrology and Earth System Sciences*, *16*(2), 551–562. https://doi.org/10.5194/hess-16-551-2012

Petheram, C., Rustomji, P., Chiew, F. H. S., & Vleeshouwer, J. (2012). Rainfall–runoff modelling in northern Australia: A guide to modelling strategies in the tropics. *Journal of Hydrology*, *462*, 28–41. https://doi.org/10.1016/j.jhydrol.2011.12.046

Pianosi, F., & Wagener, T. (2016). Understanding the time-varying importance of different uncertainty sources in hydrological modelling using global sensitivity analysis. *Hydrological Processes*, *30*(22), 3991–4003. https://doi.org/10.1002/hyp.10968

Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., & Perrin, C. (2017). Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. *Water Resources Research*, *53*(8), 7247–7268. https://doi.org/10.1002/2016wr019991

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pool, S., Viviroli, D., & Seibert, J. (2019). Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States. *Water Resources Research*, *55*(1), 363–377. https://doi.org/10.1029/2018wr023855

Post, D. A. (2009). Regionalizing rainfall–runoff model parameters to predict the daily streamflow of ungauged catchments in the dry tropics. *Hydrology Research*, *40*(5), 433–444. https://doi.org/10.2166/nh.2009.036

Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(asce)he.1943-5584.0000690

Razavi, T., & Coulibaly, P. (2016). Improving streamflow estimation in ungauged basins using a multi-modelling approach. *Hydrological Sciences Journal*, *61*(15), 2668–2679. https://doi.org/10.1080/02626667.2016.1154558

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andréassian, V., & Oudin, L. (2016). How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. *Water Resources Research*, *52*(6), 4765–4784. https://doi.org/10.1002/2015WR018549

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

Schaller, M. F., & Fan, Y. (2009). River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. *Journal of Geophysical Research*, *114*(D4), D04103. https://doi.org/10.1029/2008jd010636

Seibert, J. (1999). Regionalisation of parameters for a conceptual rainfall-runoff model. *Agricultural and Forest Meteorology*, *98*, 279–293. https://doi.org/10.1016/s0168-1923(99)00105-7

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: How many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892. https://doi.org/10.5194/hess-13-883-2009

Seibert, J., & Vis, M. J. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., Vis, M. J., Lewis, E., & Meerveld, H. V. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Singh, R., Archfield, S. A., & Wagener, T. (2014). Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments–A comparative hydrology approach. *Journal of Hydrology*, *517*, 985–996. https://doi.org/10.1016/j.jhydrol.2014.06.030

Skaugen, T., Peerebom, I. O., & Nilsson, A. (2015). Use of a parsimonious rainfall–run-off model for predicting hydrological response in ungauged basins. *Hydrological Processes*, *29*(8), 1999–2013. https://doi.org/10.1002/hyp.10315

Song, J. H., Her, Y., Suh, K., Kang, M. S., & Kim, H. (2019). Regionalization of a rainfall-runoff model: Limitations and potentials. *Water*, *11*(11), 2257. https://doi.org/10.3390/w11112257

Swain, J. B., & Patra, K. C. (2017). Streamflow estimation in ungauged catchments using regionalization techniques. *Journal of Hydrology*, *554*, 420–433. https://doi.org/10.1016/j.jhydrol.2017.08.054

Swain, J. B., & Patra, K. C. (2019). Impact of catchment classification on streamflow regionalization in ungauged catchments. *SN Applied Sciences*, *1*(5), 1–14. https://doi.org/10.1007/s42452-019-0476-6

USGS (U.S. Geological Survey). (2014). EflowStats R-package. Retrieved from https://github.com/USGS-R/EflowStats

USGS (U.S. Geological Survey). (2020). *Watershed boundary dataset for HUC#2*. Retrieved from https://developers.google.com/earth-engine/datasets/catalog/USGS_WBD_2017_HUC02#description

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–Part 3: Runoff signatures in Austria. *Hydrology and Earth System Sciences*, *17*(6), 2263–2279. https://doi.org/10.5194/hess-17-2263-2013

Viviroli, D., & Seibert, J. (2015). Can a regionalized model parameterisation be improved with a limited number of runoff measurements? *Journal of Hydrology*, *529*, 49–61. https://doi.org/10.1016/j.jhydrol.2015.07.009

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, *1*(4), 901–931. https://doi.org/10.1111/j.1749-8198.2007.00039.x

Wallace, J. M., & Hobbs, P. V. (2006). Atmospheric science: An introductory survey (second edition). In R. Dmowksa, D. Hartmann, & H. T. Rossby (Eds.), *International geophysics series*. Canada: Academic Press.

Waseem, M., Ajmal, M., & Kim, T. W. (2015). Ensemble hydrological prediction of streamflow percentile at ungauged basins in Pakistan. *Journal of Hydrology*, *525*, 130–137. https://doi.org/10.1016/j.jhydrol.2015.03.042

Yang, X., Magnusson, J., Huang, S., Beldring, S., & Xu, C. Y. (2020). 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

Yang, X., Magnusson, J., Rizzi, J., & Xu, C. Y. (2018). Runoff prediction in ungauged catchments in Norway: Comparison of regionalization approaches. *Hydrology Research*, *49*(2), 487–505. https://doi.org/10.2166/nh.2017.071

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9), W09417. https://doi.org/10.1029/2007wr006716

Zelelew, M. B., & Alfredsen, K. (2013). Sensitivity-guided evaluation of the HBV hydrological model parameterization. *Journal of Hydroinformatics*, *15*(3), 967–990. https://doi.org/10.2166/hydro.2012.011

Zhang, Y., & Chiew, F. H. (2009). Relative merits of different methods for runoff predictions in ungauged catchments. *Water Resources Research*, *45*(7), W07412. https://doi.org/10.1029/2008wr007504