Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments

Hylke E. Beck1, Ming Pan3, Peirong Lin3, Jan Seibert2,3,4, Albert I. J. M. van Dijk5, and Eric F. Wood1

1Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA, 2Department of Geography, University of Zurich, Zurich, Switzerland, 3Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, 4Department of Physical Geography, Stockholm University, Stockholm, Sweden, 5Fenner School for Environment and Society, Australian National University, Canberra, ACT, Australia

Abstract All hydrological models need to be calibrated to obtain satisfactory streamflow simulations. Here we present a novel parameter regionalization approach that involves the optimization of transfer equations linking model parameters to climate and landscape characteristics. The optimization was performed in a fully spatially distributed fashion at high resolution (0.05°), instead of at lumped catchment scale, using an unprecedented database of daily observed streamflow from 4,229 headwater catchments (<5,000 km²) worldwide. The optimized equations were subsequently applied globally to produce parameter maps for the entire land surface including ungauged regions. The approach was evaluated using the Kling-Gupta efficiency (KGE) and a gridded version of the hydrological model HBV. Tenfold cross validation was used to evaluate the generalizability of the approach and to obtain an ensemble of parameter maps. For the 4,229 independent validation catchments, the regionalized parameters yielded a median KGE of 0.46. The median KGE improvement (relative to uncalibrated parameters) was 0.29, and improvements were obtained for 88% of the independent validation catchments. These scores compare favorably to those from previous large catchment sample studies. The degree of performance improvement due to the regionalized parameters did not depend on climate or topography. Substantial improvements were obtained even for independent validation catchments located far from the catchments used for optimization, underscoring the value of the derived parameters for poorly gauged regions. The regionalized parameters—available via www.gloh2o.org/hbv—should be useful for hydrological applications requiring accurate streamflow simulations.

1. Introduction

All hydrological models, whether physical or conceptual, need to be calibrated to obtain satisfactory streamflow simulations, due to (i) the impossibility of measuring all required model parameters at the model application scale, and (ii) the simplification and spatiotemporal discretization of complex, highly heterogeneous rainfall-runoff processes (Beven, 1989; Blöschl & Sivapalan, 1995; Duan et al., 2001, 2006; McDonnell et al., 2007; Refsgaard, 1997; Vereecken et al., 2019). Conventional calibration approaches aim to improve the correspondence between observed and simulated streamflow by adjusting the model parameters either manually (e.g., Bajracharya et al., 2017; Mishra et al., 2018) or automatically (e.g., Hirpa et al., 2018; Nijssen et al., 2001; for overviews, see Gupta et al., 2013; Moradkhani & Sorooshian, 2009; Refsgaard, 1997; Yilmaz et al., 2010). These approaches typically result in uniform parameter values for each catchment and spatial discontinuities between adjacent catchments. Additionally, they are not applicable in ungauged regions, which comprise the large majority of the global land surface (Fekete & Vörösmarty, 2007; Hannah et al., 2011; Sivapalan, 2003).

Regionalization approaches are designed to estimate model parameters in ungauged regions through the transfer of knowledge from gauged to ungauged catchments (see reviews by Beck et al., 2016; Blöschl et al., 2013; Hrachowitz et al., 2013; Parajka et al., 2013; Razavi & Coulthab, 2013; Samaniego et al., 2010). The most widely used regionalization approaches involve catchment-by-catchment calibration followed by (i) regression with landscape and climate predictors (e.g., Abdulla & Lettenmaier, 1997; Döll et al., 2003;
Livneh & Lettenmaier, 2013), (ii) transfer of parameter sets based on geographic proximity (e.g., Merz & Blöschl, 2004; Widén-Nilsson et al., 2007), or (iii) transfer of parameter sets based on physical or climatic similarity (e.g., Beck et al., 2016; Nijssen et al., 2001). However, these approaches assume lumped (i.e., spatially uniform) parameter values for each catchment and thus neglect the often pronounced within-catchment heterogeneity in landscape and climate (Kling & Gupta, 2009; Rouholahnejad-Freund et al., 2019). Additionally, they ignore the discrepancy in scale and thus rainfall-runoff behavior between catchments and grid cells (Becker & Braun, 1999; Blöschl & Sivapalan, 1995). Furthermore, the regression method is confounded by parameter equifinality (i.e., different parameter sets yielding the same results Beven, 2006; Kokkonen et al., 2003), while the geographic-proximity method should only be used in data-rich regions (Oudin et al., 2008; Reichl et al., 2009; Vandewiele & Elias, 1995). Another type of regionalization approach employs observation-based maps of streamflow signatures (such as baseflow index) to constrain model parameters (e.g., Boughton & Chiew, 2007; Troy et al., 2008; Yadav et al., 2007). However, this approach is confounded by the limited predictability and information content of streamflow signatures (Addor et al., 2018; Beck et al., 2015). Yet another type involves the simultaneous optimization of model parameters for catchments grouped based on landscape and climate (e.g., Arheimer et al., 2019; Huang et al., 2019), but this approach fails to account for the within-catchment heterogeneity in landscape and climate and yields uniform parameters for each group.

A different regionalization approach that overcomes most of the aforementioned limitations involves the optimization of coefficients of transfer equations linking model parameters to landscape and climate predictors. Notable examples include Hundecha and Bárdossy (2004), who calibrated the HBV model for 95 catchments in the European Rhine basin; Bastola et al. (2008), who calibrated TOPMODEL for 26 catchments in the UK; Rakovec et al. (2016), who calibrated the mHM model for 400 catchments across Europe; and Mizukami et al. (2017), who calibrated the VIC model for 531 catchments in the conterminous United States. A limitation of these studies, however, is their regional focus and therefore limited generalizability. Additionally, three of these studies (Hundecha & Bárdossy, 2004; Mizukami et al., 2017; Rakovec et al., 2016) did not incorporate climate-related predictors, despite several large-scale regionalization studies highlighting their value (e.g., Beck et al., 2016; Nijssen et al., 2001; Singh et al., 2014; Young, 2006). Furthermore, Hundecha and Bárdossy (2004) and Bastola et al. (2008) used lumped predictor and parameter values for the catchments and thus neglected the within-catchment heterogeneity in hydrological processes (Kling & Gupta, 2009; Rouholahnejad-Freund et al., 2019; Samaniego et al., 2017). Rakovec et al. (2016) and Mizukami et al. (2017) ran their hydrological models in a spatially distributed fashion and thus did account for within-catchment heterogeneity, although they used fairly coarse spatial grids (0.25° and 0.125°, respectively).

Here we present a novel parameter regionalization approach using transfer equations to derive an ensemble of parameter maps for the global land surface including ungauged regions. The transfer equations were optimized in a spatially distributed fashion at high resolution (0.05°) based on an unprecedented database of daily observed streamflow from 4,229 headwater catchments (<5,000 km²) worldwide. The approach was implemented using a gridded version of the HBV hydrological model forced with state-of-the-art downscaled meteorological data. Tenfold cross validation was used to evaluate the generalizability of the approach and to quantify uncertainty globally. We address the following aspects of the approach: (i) the model performance, (ii) the factors determining the performance, and (iii) the spatial patterns of the regionalized parameters.

2. Data and Methods

2.1. Regionalization Approach

Our regionalization approach involves the optimization of coefficients in transfer equations linking model parameters to predictors related to climate, land cover, topography, and soils (Figure 1). For the optimization, the hydrological model was run at a daily temporal and a 0.05° spatial resolution, and the runoff outputs were spatially aggregated for each catchment. The aggregated runoff was subsequently compared to the observed streamflow of each catchment through the computation of a performance score, after which a mean performance score was calculated over all catchments. We only used observed streamflow from catchments up to 5,000 km², due to the dominance of channel routing effects in larger catchments at the daily

2.1. Regionalization Approach

Our regionalization approach involves the optimization of coefficients in transfer equations linking model parameters to predictors related to climate, land cover, topography, and soils (Figure 1). For the optimization, the hydrological model was run at a daily temporal and a 0.05° spatial resolution, and the runoff outputs were spatially aggregated for each catchment. The aggregated runoff was subsequently compared to the observed streamflow of each catchment through the computation of a performance score, after which a mean performance score was calculated over all catchments. We only used observed streamflow from catchments up to 5,000 km², due to the dominance of channel routing effects in larger catchments at the daily
The model was only run for the 171,128 grid cells which were included in any of the catchments (comprising 2.2% of the entire land surface), to use computational resources efficiently.

The transfer equations established for the model parameters have the following form:

$$MP_i = a_{i1} HI + a_{i2} MAP + a_{i3} PET + a_{i4} NDVI + a_{i5} OW + a_{i6} Slope + a_{i7} Sand + a_{i8} Clay + a_{i9},$$  \hspace{1cm} (1)

where $MP_i$ is the $i$th model parameter; $a_{i1−9}$ are the coefficients that need to be optimized; and HI, MAP, PET, NDVI, OW, Slope, Sand, and Clay are the clipped and standardized predictors (defined in Table 1).

We only considered linear relationships with the predictors (except for MAP), to avoid highly skewed predictor distributions which might lead to overfitting. Although this might not be ideal for every parameter-predictor combination (Heuvelmans et al., 2006; Samaniego et al., 2010), determining the most appropriate shape for each parameter-predictor relationship is beyond the scope of the current study.
Among the eight predictors, three were related to climate (AI, MAP, and PET), two to land cover (NDVI and OW), one to topography (Slope), and two to soils (Sand and Clay). Climate-related predictors were used because climate is known to influence vegetation, soils, and geomorphology and thus exerts a major indirect influence on the rainfall-runoff response (Gentine et al., 2012; Troch et al., 2013). Additionally, several continental- and global-scale regionalization studies have highlighted the value of incorporating climate-related predictors (e.g., Beck et al., 2016; Nijssen et al., 2001; Singh et al., 2014; Young, 2006). The MAP predictor was square root transformed to render the data more normally distributed. The NDVI predictor was added because of its predictive power in previous regionalization studies (Beck, van Dijk, et al., 2013; Van der Velde et al., 2013). The Sand and Clay predictors were used because soil texture has a strong influence on rainfall-runoff processes (Hewlett, 1961; Price, 2011; Zecharias & Brutsaert, 1988) and is correlated with several streamflow signatures (Beck, van Dijk, et al., 2013; Santhi et al., 2008). The Slope predictor was included because surface slope is correlated with soil depth (Tesfa et al., 2009) and because more steeply sloping aquifers are expected to drain faster (Brutsaert & Nieber, 1977; Vogel & Kroll, 1996; Zecharias & Brutsaert, 1988).

Each predictor was preprocessed as follows. First, we up-scaled the data to 0.05° using bilinear averaging. Second, we filled any gaps using nearest neighbor. Third, we clipped the values using the 1st and 99th percentiles of the area covered by the catchments, to avoid application of the transfer equations outside the range of the catchment distribution of predictor values. Fourth and last, we standardized the values by subtracting the mean and dividing by the standard deviation of the area covered by the catchments, to make all predictors intercomparable. Since all predictor fields have a native resolution ≤1 km (Table 1), the optimized transfer equations can be used to derive parameter maps at up to 1-km resolution globally.

### 2.2. Observed Streamflow and Catchment Selection

We used an initial database of daily observed streamflow and catchment boundaries for 21,955 stations worldwide (Beck et al., 2020). The database was compiled from seven different national and international sources (listed in descending order of the number of catchments): (i) the United States Geological Survey (USGS) National Water Information System (NWIS; https://waterdata.usgs.gov/nwis) and GAGES-II database (Falcone et al., 2010; 9,180 catchments); (ii) the Global Runoff Data Centre (GRDC; https://grdc.bafg.de; Lehner, 2012; 4,628 catchments); (iii) the HidroWeb portal of the Brazilian Agência Nacional de Águas (https://www.snirh.gov.br/hidroweb; 3,029 catchments); (iv) the European Water Archive (EWA) of EURO-FRIEND-Water (https://ne-friend.bafg.de) and the CCM2-JRC CCM River and Catchment Database (https://inspire-geoportal.ec.europa.eu/demos/ccm; Vogt et al., 2007; 2,260 catchments); (v) Water Survey of Canada (WSC) National Water Data Archive (HYDAT; https://https://waterdata.usgs.gov/nwis) and GAGES-II database (Falcone et al., 2010; 9,180 catchments); (vi) the Australian Bureau of Meteorology (BoM; https://www.bom.gov.au/waterdata; Zhang et al., 2013; 776 catchments) and (vii) the Chilean Center for Climate and Resilience Research (CR2) website (https://www.cr2.cl/recursos-y-publicaciones/bases-de-datos/datos-de-caudales) and the CAMELS-CL data set (Alvarez-Garreton et al., 2018; 531 catchments).

Unsuitable catchments were identified using the following criteria. First, GRDC catchments without daily streamflow data were discarded. Second, “nonreference” GAGES-II catchments were discarded. Third, catchments smaller than 50 km² were discarded, to ensure that each catchment covers at least two 0.05° grid cells. Fourth, catchments larger than 5,000 km² were discarded, as channel routing effects become apparent at the daily time scale in larger catchments (Gericke & Smithers, 2014). Fifth, catchments with less than 5 yr of streamflow data during 2000–2016 (the calibration period) were discarded. Sixth, we discarded catchments with potentially erroneous streamflow data (identified through visual screening). The final database used for the regionalization comprised 4,229 catchments (median size 695 km²; Figure 2).

### 2.3. Climate Groups and Cross Validation

To reduce the computational load, minimize the possibility of underfitting, and reduce the heterogeneity among catchments, we subdivided the land surface and the catchment set into three groups based on the Köppen-Geiger (KG) classification (Beck et al., 2018): (i) tropical (Class A; 566 catchments); (ii) arid and temperate (Classes B and C, respectively; 1,802 catchments); and (iii) cold and polar (Classes D and E,
respectively; 1,861 catchments; Figure 2). The arid and polar classes were represented by a small number of catchments (122 and 150, respectively) and were therefore merged with the temperate and cold classes, respectively. The dominant class of each catchment was used to subdivide the catchment set.

The transfer equation coefficients (Equation 1) were optimized for each of the three climate groups separately. For the optimization, we used tenfold cross validation, allowing us to (i) estimate the generalizability of the derived parameters, and (ii) obtain an ensemble of parameter maps, the spread of which provides an indication of the uncertainty. For each cross-validation iteration, the catchment set was partitioned into subsets of 90% for calibration and 10% for validation. The partitioning was random with each catchment being used only once for validation.

2.4. Hydrological Model and Meteorological Forcing

While the parameter regionalization approach can be applied to any hydrological model, we tested it using a gridded implementation of the HBV model (Bergström, 1976, 1992; Seibert & Vis, 2012). HBV was used because of its low complexity, high agility, and computational efficiency. Additionally, the model has been successfully used in numerous studies spanning a wide range of climate and physiographic conditions (e.g., Beck, Bruijnzeel, et al. 2013; Breuer et al., 2009; Deelstra et al., 2010; Demirel et al., 2015; Driessen et al., 2010; Jódar et al., 2018; Pesca et al., 2012; Steele-Dunne et al., 2008; Te Linde et al., 2008; Vetter et al., 2015), including several parameter regionalization studies (e.g., Bárdossy, 2007; Beck et al., 2016; Booij, 2005; Jin et al., 2009; Hundecha & Bárdossy, 2004; Masih et al., 2010; Merz & Blöschl, 2004; Parajka et al., 2005, 2007; Seibert, 1999). The model runs at a daily time step, has one unsaturated zone store, two groundwater stores, and 12 free parameters (Table 2).

Transfer equations were established for all 12 model parameters, resulting in \(12 \times 9 = 108\) coefficients that required optimization (Equation 1). We thus calibrated all model parameters and linked each to all eight predictors, to maximize the model agility (Mendoza et al., 2015). This was computationally feasible and did not lead to overfitting issues (i.e., low validation scores compared to calibration) because of the large sample of catchments used. We do not expect that all coefficients are well constrained after the optimization, as some model parameters might be “insensitive” (i.e., have little influence on the simulated runoff). See Abebe et al. (2010) and Zelelew and Alfredsen (2012) for sensitivity analyses of HBV.

The model requires daily time series of precipitation, potential evaporation, and air temperature as input. For precipitation, we used the gauge-, satellite-, and reanalysis-based MSWEP data set (V2.2; 0.1° resolution; 1979–present; www.gloh2o.org/mswep; Beck, van Dijk, Levizzani, et al., 2017; Beck, Wood, et al., 2019). MSWEP was chosen for its superior performance in numerous precipitation data set evaluation studies (e.g., Alijanian et al., 2017; Bai & Liu, 2018; Beck, Pan, et al., 2019; Beck, Vergopolan, et al., 2017; Casson et al., 2018; Sahlu et al., 2017; Satgé et al., 2019; Zhang et al., 2019). The precipitation data were corrected for gauge

![Figure 2. The dominant Köppen-Geiger climate classes of the 4,229 catchments.](attachment:image-url)
undercatch and orographic effects using the PBCOR WorldClim V2 data set (Beck et al., 2020; www.gloh2o.org/pbcor) and downscaled to 0.05° using nearest neighbor resampling. Potential evaporation was estimated using the Hargreaves (1994) equation from daily minimum and maximum air temperature. Temperature estimates were obtained by averaging two reanalysis data sets, ERA-Interim (−0.75° resolution; 1979–present Dee et al., 2011; https://cds.climate.copernicus.eu) and JRA-55 (−0.56° resolution; 1959–present Kobayashi et al., 2015; https://jra.kishou.go.jp). Prior to the averaging, both data sets were downscaled to 0.05° and bias corrected on a monthly basis through an additive approach using the comprehensive station-based WorldClim V2 climatology (1-km resolution Fick & Hijmans, 2017; www.worldclim.org).

We only ran the model for the period 1990–2016 rather than for the entire period of forcing data availability (1979–2016), to reduce the computational demand and to take advantage of the better quality of the precipitation data after 2000 (Beck, Wood, et al., 2019); the first 10 yr of the record (1990–1999) were used only to initialize the model stores.

### 2.5. Performance Metric

For each catchment, we calculated scores of a transformed version of the Kling-Gupta efficiency (KGE$_o$) from daily time series of observed streamflow and simulated streamflow (obtained by aggregating the simulated runoff, Figure 1) for the period 2000–2016. The original (untransformed) KGE is an objective performance metric combining correlation, bias, and variability introduced by Gupta et al. (2009) and modified by Kling et al. (2012). It is defined according to

$$KGE = 1 - \sqrt{(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2},$$

where the correlation component $r$ is represented by Pearson’s correlation coefficient, the bias component $\beta$ by the ratio of estimated and observed means, and the variability component $\gamma$ by the ratio of the estimated and observed coefficients of variation:

$$\beta = \frac{\mu_e}{\mu_o} \quad \text{and} \quad \gamma = \frac{\sigma_e}{\mu_o}/\frac{\sigma_o}{\mu_o},$$

where $\mu$ and $\sigma$ represent the distribution mean and standard deviation, respectively, and the subscripts $e$ and $o$ indicate estimate and reference, respectively. A drawback of the KGE is that it does not have a lower limit. Accordingly, the mean over a large sample of catchments can be dominated by a single catchment with an extremely low KGE value. To avoid this, we applied the following transformation (Mathevet et al., 2006):

$$KGE_o = \frac{KGE}{2 - KGE}.$$  

The resulting KGE$_o$ values range from $-1$ to $1$, with higher values indicating better performance.

### 2.6. Optimization Algorithm

For all three climate group and all 10 cross-validation iterations, we optimized the coefficients of the transfer equations using the $(\mu + \lambda)$ evolutionary algorithm implemented using the Distributed Evolutionary Algorithms in Python (DEAP) toolkit (Fortin et al., 2012). The population size ($\mu$) was set at 16 and the recombination pool size ($\lambda$) at 32. Each generation produced $\lambda$ offspring from the population. Offspring were evaluated after which the population of the next generation was selected from both offspring and population. Crossover and mutation probabilities were set at 0.9 and 0.1, respectively. The number of generations was set at 25, as this was found to be sufficient for achieving convergence. This resulted in $3 \times 10 \times 25 \times 32 = 24,000$ gridded model evaluations.

### 2.7. Local Calibration

To obtain an upper limit of what is feasible using the combination of meteorological forcing, observed streamflow, and catchment boundary data (Selbert et al., 2018), we calibrated HBV on a catchment-by-catchment basis against locally observed streamflow data. For this purpose, we ran the model in a lumped (i.e., nondistributed) fashion with catchment-mean meteorological forcing data for each of the 4,229 catchments. The first half of the record of simultaneous forcing and observed streamflow data (1979–2016) was used for validation and the second half for calibration. The stores were initialized by running the model
for the first 10 yr of the record if the record length was 10 yr or by running the model twice for the entire record if the record length was <10 yr. The KGE was used as objective function, and the \((\mu + \lambda)\) evolutionary algorithm was used for the optimization, with \(\mu\) set to 20, \(\lambda\) set to 40, and the number of generations set to 12 (resulting in \(4,229 \times 20 \times 40 \times 12 = 40.6\) million lumped model evaluations).

### 3. Results and Discussion
#### 3.1. Model Performance

Figure 3 presents the streamflow simulation performance obtained using uncalibrated parameters (randomly generated in the first generation of the optimization process), regionalized parameters for the independent evaluation catchments, and local calibration for the validation period. The median daily KGE using uncalibrated parameters over all catchments was 0.17 (Figure 3c), whereas the median daily KGE obtained using regionalized parameters for the calibration catchments was 0.46. The median daily KGE obtained using regionalized parameters for the independent validation catchments was also 0.46 (Figure 3b), suggesting that the approach generalizes well to catchments that were not used for the optimization and hence to ungauged regions. The difference in median daily KGE between uncalibrated parameters (0.17) and regionalized parameters (0.46) was thus 0.29, and improvements were obtained for 88% of the validation catchments (Figure 3a). These results confirm the efficacy of our regionalization approach in improving the streamflow simulation performance. Our median daily KGE using uncalibrated parameters of 0.17 (Figure 3c) exceeds the range of median daily KGE values of −0.25 to 0.13 obtained by Beck et al. (2016) for a diverse set of 10 uncalibrated models in catchments comparable to ours, suggesting that we are not overestimating the benefit of the regionalization.

The median daily KGE obtained using local calibration was 0.77 for the calibration period and 0.69 for the validation period (Figure 3d). The difference in median KGE between regionalized parameters (0.46) and locally calibrated parameters (0.69) was thus 0.23. Such a difference is to be expected, not least because scores obtained using local calibration compensate for local errors in meteorological forcing (Beck, Pan, et al., 2019; Beck, Vergopolan, et al., 2017), observed streamflow (Di Baldassarre & Montanari, 2009; McMillan et al., 2010), and catchment boundary data (Kauffeldt et al., 2013; Lehner, 2012).

Our local calibration scores (Figure 3d) were similar to or higher than those from previous studies, suggesting that our meteorological forcing data are of good quality (in agreement with several precipitation data set evaluations; e.g., Beck, Pan, et al., 2019; Beck, Vergopolan, et al., 2017) and that our model setup is sufficiently agile (i.e., capable of simulating a wide range of catchment behavior; Mendoza et al., 2015). Alfieri et al. (2020), for example, calibrated the LISFLOOD model for 1,226 catchments globally (mean size 42,000 km²) and obtained median daily KGE values of 0.67 and 0.61 for the calibration and validation periods, respectively, while we obtained higher median scores of 0.77 and 0.69, respectively (Figure 3d), despite the much smaller size of our catchments (mean size 1,165 km²). Filipova and Leedal (2018) calibrated the IHACRES model for ∼3,000 catchments globally similar in size to ours and obtained a median daily calibration KGE of 0.60, considerably lower than our median score of 0.77. For the conterminous United States, we obtained a median daily validation KGE of 0.73, whereas Mizukami et al. (2019) obtained median scores of 0.63 and 0.74 for the VIC and mHM models, respectively, and Gao et al. (2019) obtained median scores of 0.62 and 0.61 for HBV and TOPMODEL, respectively.

To our knowledge, only seven previous regionalization studies had a global scope (Arheimer et al., 2019; Beck et al., 2016; Döll et al., 2003; Filipova & Leedal, 2018; Nijssen et al., 2001; Van Dijk et al., 2013; Widén-Nilsson et al., 2007). However, none of these studies accounted for the within-catchment heterogeneity in landscape and climate (Kling & Gupta, 2009; Rouholahnejad-Freund et al., 2019) or the scale discrepancy between catchments and grid cells (Becker & Braun, 1999; Blöschl & Sivapalan, 1995). Additionally, three studies (Döll et al., 2003; Nijssen et al., 2001; Widén-Nilsson et al., 2007) only used observed streamflow from large catchments (≥10,000 km²) in which routing effects tend to dominate at the daily time scale (Gericke & Smithers, 2014), while one study (Widén-Nilsson et al., 2007) used a regionalization approach based on spatial proximity despite a lack of gauges in many regions across the globe (Fekete & Vörösmarty, 2007; Hannah et al., 2011; Sivapalan, 2003). Furthermore, with the exception of Beck et al. (2016), Filipova and Leedal (2018), and Arheimer et al. (2019), these studies used observed
streamflow from a relatively small number of catchments (9 to 311) and did not evaluate the performance of the regionalized parameters in independent validation catchments.

In probably the most similar previous global regionalization study to date, Beck et al. (2016) produced parameter maps (0.5° resolution) for HBV using a regionalization approach based on climatic and physiographic similarity. Although they used another performance metric to optimize their parameters, they reported a median daily KGE improvement using regionalized parameters of 0.08, considerably lower than our median improvement of 0.29 (Figure 3a). Furthermore, they reported performance improvements for 79% of the validation catchments using the performance metric that they used for the optimization, substantially less than the 88% obtained by us (Figure 3a). This suggests that our new regionalization approach provides a better generalization capability.

Filipova and Leedal (2018) used an approach based on climatic and physiographic similarity to regionalize parameters of the IHACRES model for ~3,000 catchments globally similar in size to ours. They obtained a median daily KGE of 0.40 using leave-one-out cross validation, which is lower than our median score of 0.46 (Figure 3b).

Figure 3. (a) The improvement in KGE after regionalization calculated as the difference between (b) KGE obtained for the independent validation catchments using regionalized parameters and (c) KGE obtained for the first generation of the optimization process (i.e., using uncalibrated parameters). (d) KGE obtained for the validation period using parameters calibrated against locally observed streamflow. Each data point represents a catchment centroid (N = 4,229).
Arheimer et al. (2019) produced global parameter maps for the HYPE hydrological model by jointly optimizing the parameters of groups of catchments. They obtained a median monthly KGE of 0.40 for 2,863 large catchments (≫1,000 km²), whereas we obtained a median daily KGE of 0.46 for 4,229 small-to-medium-sized catchments (median area 695 km²; Figure 3b). Thus, their median score is lower than ours, despite the fact that simulation performance might be expected to be better at monthly resolution (Beck et al., 2016; Xia et al., 2012; Zink et al., 2017) and in large catchments, where (i) local forcing errors and landscape heterogeneity may be averaged out, (ii) the hydrograph is smoother due to channel routing, and (iii) the catchment delineation is likely more accurate (Merz et al., 2009; Parajka et al., 2013; Rakovec et al., 2016).

3.2. Factors Determining the Performance

The improvement in KGE using regionalized parameters relative to uncalibrated parameters showed no clear relationship with catchment-mean humidity index (Figure 4a) or topographic slope (Figure 4b), suggesting that the benefit of the regionalized parameters does not depend on climate or topography. Conversely, there was a weak positive relationship between KGE improvement values and catchment area (Figure 4c), suggesting that larger catchments benefit more from the regionalization approach. This could be because larger catchments aggregate the runoff over multiple grid cells, potentially canceling out the parameter and forcing errors present in individual grid cells. In a similar vein, several previous studies obtained better performance by averaging the outputs from multiple parameter sets than by taking the output of a single parameter set (e.g., Bao et al., 2012; Beck et al., 2016; Garambois et al., 2015; McIntyre et al., 2004, 2005; Oudin et al., 2008; Reichl et al., 2009; Viney et al., 2009; Zhang & Chiew, 2009). A weak negative relationship was found between KGE improvement values and “distance” (defined as the mean distance to the 10 closest catchments used for the optimization; Figure 4d), suggesting that the regionalization approach provides slightly less (but still substantial) benefit in poorly gauged regions.

We found a strong positive correlation (Spearman’s rank correlation coefficient $\rho = 0.39$) between regionalization KGE values and catchment-mean humidity index (Figure 4e), indicating that model performance tends to be better in humid regions, confirming numerous studies (e.g., Arheimer et al., 2019; Beck et al., 2016; Beck, van Dijk, et al., 2017; Beck, Vergopolan, et al., 2017; Essou et al., 2016; Newman et al., 2015; Parajka et al., 2013) and previously attributed to the localized short-lived convective rainfall, the high evaporative losses, and the nonlinear rainfall-runoff relationship (Pilgrim et al., 1988; Ye et al., 1997). Conversely, regionalization KGE values were not clearly related to catchment-mean topographic slope (Figure 4f). The slightly negative relationship between regionalization KGE values and catchment area...
Figure 4g is contrary to previous studies (e.g., Gericke & Smithers, 2014; Parajka et al., 2013) and is probably partly due to the more arid nature of the larger catchments in our data set ($\rho = -0.19$ between catchment area and catchment-mean humidity index), and partly because we did not use a routing model to account for streamflow delays in larger catchments. Similarly, the negative relationship between regionalization KGE values and distance (Figure 4h) likely reflects the fact that arid regions tend to be more poorly gauged ($\rho = -0.17$ between distance and catchment-mean humidity index).

Among the five Köppen-Geiger climate classes, the lowest KGE values were obtained for the arid class, and the highest for the temperate and cold classes (Figure 5), similar to previous large catchment sample studies (Beck et al., 2016; Beck, van Dijk, et al., 2017; Beck, Vergopolan, et al., 2017). The good performance in temperate regions is attributable to the relative simplicity of the hydrological response and the dense precipitation measurement network (Kidd et al., 2017; Schneider et al., 2014), while in cold regions it is likely attributable to the smoothly varying seasonal cycle of streamflow and the predictability of frontal weather systems (Beck, Vergopolan, et al., 2017; Ebert et al., 2007). The somewhat greater spread among KGE values for the tropical class (Figure 5) may reflect the large variability in precipitation measurement network density across the tropics (Kidd et al., 2017; Schneider et al., 2014). The slightly greater spread among regionalization KGE values than among uncalibrated KGE values (Figure 5) suggests that catchments for which the simulation performance was poor prior to model parameter optimization (e.g., due to model structural or forcing data deficiencies) derived less benefit overall from our regionalization approach. The substantial spread among KGE values obtained using local calibration for the polar class (Figure 5) likely reflects the inability of lumped models to simulate streamflow in mountainous, snowmelt-dominated catchments.

3.3. Spatial Patterns of the Regionalized Parameters

Figure 6 presents maps of four key HBV model parameters (BETA, FC, K2, and PERC; Table 2) derived using the new regionalization approach. Our maps vary according to landscape and climate characteristics for the entire global land surface including ungauged regions. Although many studies have optimized HBV parameters (e.g., Bárdossy, 2007; Beck et al., 2016; Booij, 2005; Hundecha & Bárdossy, 2004; Jin et al., 2009; Masih et al., 2010; Merz & Blöschl, 2004; Parajka et al., 2005, 2007; Seibert, 1999), only two have actually published maps of their optimized parameters to which we can compare our regionalized parameter maps (Beck et al., 2016; Merz & Blöschl, 2004). The maps of Merz and Blöschl (2004, their Figures 4–7), derived by calibrating HBV for 308 Austrian catchments individually, exhibit reasonable agreement with ours (Figure 6): For example, both exhibit lower FC (maximum soil moisture storage) in the northeastern lowlands of Austria and higher BETA (shape coefficient of recharge function) in the mountainous west. The maps of Beck et al. (2016, their Figure 4), derived by optimizing parameters for each catchment individually and transferring them to “similar” 0.5° grid cells globally, also exhibit reasonable agreement with ours.
However, both studies optimized the parameters for the catchments individually, and therefore their parameter maps lack spatial coherence. Conversely, we optimized transfer equation coefficients in a spatially distributed fashion at high resolution for large groups of catchments jointly, resulting in parameter maps with high spatial coherence.

However, it can be difficult to explain the spatial patterns of the regionalized parameters, due to identifiability issues (Sorooshian & Gupta, 1983) and parameter interactions (Gupta & Sorooshian, 1983). Yet, the higher BETA in arid regions (Figure 6a) reduces the recharge rate to the groundwater stores, which is in line with global recharge assessments based on conceptual (Döll & Fiedler, 2008) and statistical (Mohan et al., 2018) models. The patterns in K2 (recession coefficient of lower zone; Figure 6c) can be explained by the sand content of the soil (Hengl et al., 2017), which is consistent with the notion that the permeability of the soil has a strong impact on baseflow (Hewlett, 1961; Price, 2011; Zecharias & Brutsaert, 1988). The lower PERC (maximum percolation to lower zone) in arid regions (Figure 6d) reduces the flow of water from the upper groundwater store (which has a quick outflow) to the lower groundwater store (which has a slow outflow) and hence results in more rapidly receding runoff, consistent with baseflow recession assessments using observed streamflow for the pantropics (Peña-Arancibia et al., 2010) and the globe (Beck et al., 2015; Beck, van Dijk, et al., 2013). The patterns in FC (Figure 6b) are less easily interpretable; the low FC in tropical regions is counterintuitive given the thick regolith cover (Bonell, 2005). It may reflect parameter compensation behavior within the model to reduce evaporation and increase runoff.

## 4. Conclusion

Previous regionalization approaches generally ignored the within-catchment variability in climate and landscape, did not optimize the model parameters for all catchments jointly, failed to evaluate the performance in independent validation catchments, and had a regional focus and thus questionable generalizability. To overcome these limitations, we introduced a novel regionalization approach and implemented it for a gridded version of the HBV model using a uniquely large database of daily streamflow data for 4,229 catchments worldwide. An ensemble of high-resolution (0.05°) parameter maps covering the entire land surface including ungauged regions was derived. Our findings can be summarized as follows:

1. The regionalized parameters yielded, for the independent validation catchments, a median daily KGE of 0.46. The median KGE improvement due to the regionalization (relative to uncalibrated parameters) for the independent validation catchments was 0.29, with improvements obtained for 88% of the independent validation catchments. Our scores compare favorably to those from previous continental- and global-scale studies, confirming the effectiveness of our approach in improving streamflow simulation performance.
2. The performance improvement due to the regionalized parameters did not depend on climate or topography. Substantial improvements were obtained even for independent validation catchments located far away from the catchments used for optimization, highlighting the value of the approach for poorly gauged regions. The streamflow simulation performance was worst in arid regions and best in temperate and cold regions, in agreement with several previous large catchment sample studies. The spread in performance was greatest among tropical catchments.

3. In contrast to conventional catchment-by-catchment calibration approaches, the new regionalization approach yields parameters that vary according to landscape and climate characteristics for the entire land surface including ungauged regions. The obtained parameter maps exhibit reasonable agreement with those from two previous studies. We were able to interpret the spatial patterns of most of the regionalized parameters based on hydrological process understanding.

Data Availability Statement

The HBV parameter maps and the optimized transfer equation coefficients can be downloaded online (via www.gloh2o.org/hbv). The majority of the streamflow, meteorological forcing, and predictor data can be obtained via the URLs listed in sections 2.2 and 2.4 and Table 1, respectively.

Acknowledgments

The following organizations are thanked for providing observed streamflow data: the United States Geological Survey (USGS), the Global Runoff Data Centre (GRDC), the Brazilian Agência Nacional de Águas, EURO-FRIEND-Water, the Water Survey of Canada (WSC), the Australian Bureau of Meteorology (BoM), and the Chilean Chilean Center for Climate and Resilience Research (CR2). We are thankful to Ross Woods for helpful comments on an earlier draft of this paper. The editor and two of the three anonymous reviewers are thanked for constructive comments which helped us to improve the paper. Hylke Beck was supported in part by the U.S. Army Corps of Engineers’ International Center for Integrated Water Resources Management (ICWaRM), under the auspices of UNESCO.

References


