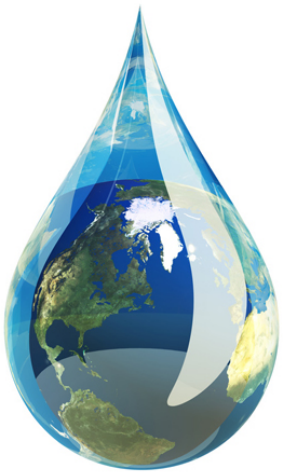
Chair of Hydrological Modeling and Water Resources

Albert-Ludwigs-University of Freiburg

Robin Schwemmle

Climatic and Physiographic Controls on Errors of Large-scale Hydrological Models



MSc-Thesis under the guidance of JProf. Dr. Andreas Hartmann

Freiburg i. Br., August 2018

Chair of Hydrological Modeling and Water Resources

Albert-Ludwigs-University of Freiburg

Robin Schwemmle

Climatic and Physiographic Controls on Errors of Large-scale Hydrological Models

Examiner: JProf. Dr. Andreas Hartmann

Co-Examiner: Dr. Ingeborg de Graaf

MSc-Thesis under the guidance of JProf. Dr. Andreas Hartmann

Freiburg i. Br., August 2018

Declaration of Authorship

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Freiburg i. Br., 30 August 2018

Robin Schwemmle

"Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand."

– Albert Einstein

Table of Contents

List of Figures

List of Tables

List of Figures in Appendix

List of Tables in Appendix

List of Abbreviations

ANN	Artificial Neural Network
DGVM	Dynamic Global Vegetation Model
EDAS	Eta Data Assimilation System
ev	Evaporites
EXP	exponential
FHLU	Fundamental Hydrologic Landscape Unit
GCM	Global Circulation Model
GHM	Global Hydrological Model
GLDAS	Global Land Data Assimilation System
GLiM	Global Lithological Map
GRDC	Global Runoff Date Centre
GSWP-2	Global Soil Wetness Project phase 2
HBV	Hydrologiska Byråns Vattenbalansavdelning
HLR	Hydrologic landscape region
HLR 1	Humid plains with permeable bedrock
HLR 10	(Sub-)Humid mountains with permeable bedrock
HLR 11	(Sub-)Humid plains/plateaus with impermeable bedrock
HLR 12	Very humid plains with impermeable bedrock
HLR 2	(Sub-)Humid plains with impermeable bedrock
HLR 3	(Sub-)Humid plains with very permeable bedrock
HLR 4	Humid low range mountains with impermeable bedrock
HLR 5	Humid plains with impermeable bedrock
HLR 6	Subhumid and (semi-)arid mountains with permeable bedrock
HLR 7	Humid mountains with impermeable bedrock

HLR 8	Subhumid plains/plateaus with permeable bedrock
HLR 9	(Sub-)Humid plains with very impermeable bedrock
ISI-MIP	Inter-Sectoral Impact Model Intercomparison Project
LC	Land cover
LHM	Large-scale Hydrological Model
LIN	linear
LOG	logarithmic
LSM	Land Surface Model
MacPDM	Macro-scale–Probability-Distributed Moisture model
MSWEP	Multi-Source Weighted-Ensemble Precipitation
mt	Metamorphic rocks
NCEP-DOE	National Centers for Environmental Prediction/Department of Energy
NetCDF	Network Common Data Format
NLDAS	North American Land Data Assimilation System
pa	Acid plutonic rocks
pb	Basic plutonic rocks
PD	Partial dependence
pi	Intermediate plutonic rocks
POW	power
py	Pyroclastics
RF	Random forest
sc	Carbonate sedimentary rocks
sm	Mixed sedimentary rocks
So	Soils
ss	Siliclastic sedimentary rocks
su	Unconsolidated sediments
T	Topography

THREDDS	Thematic Real-Time Environmental Distributed Data Services
va	Acid volcanic rocks
vb	Basic volcanic rocks
vi	Intermediate volcanic rocks
WaterGap	Water Global Assessment and Prognosis
WaterMIP	Water Model Intercomparison Project
WBM	Water Balance Model
WCI	Water Cycle Integrator
WFDEI	WATCH Forcing Data ERA-Interim
WU	Water use

List of Symbols

β	Ratio of mean simulated runoff and observed runoff	–
γ	Ratio of coefficient of variation of simulated runoff and observed runoff	–
σ	Standard deviation	–
AI	Aridity index	–
B_q	Bias at a certain flow percentile	–
B_{rel}	Relative bias	–
$B_{std-sqrt}$	Standardized-square-rooted bias	–
B_{std}	Standardized bias	–
BFI	Base flow index	–
$CLAY$	Soil clay content	%
$CORR$	Seasonal correlation between water supply and demand	–
CV	Coefficient of variation	–
$ELEV$	Surface slope	$m\ MSL$
fLi_{xx}	Fraction of lithologic class	–
fS	Fraction of snow cover	–
fW	Fraction covered by lakes and reservoirs	–
$GLAC$	Fraction covered by glaciers	–
IRR	Percentage of irrigated area	%
KGE	Kling-Gupta-Efficiency	–
$KGE_{\gamma\beta}$	Modified Kling-Gupta-Efficiency without the r term	–
$NDVI$	Normalized difference vegetation index	–
NSE	Nash-Sutcliffe-Efficiency	–
P	Mean annual potential evaporation	$mm\ yr^{-1}$
P	Mean annual precipitation	$mm\ yr^{-1}$
P_{si}	Precipitation seasonality	–
$PERM$	Permeability of geology	$log_{10}\ m^2$
PET_{si}	Potential evaporation seasonality	–
PF	Permafrost abundance	–
Q	Runoff	$mm\ day^{-1}$

Q_{obs}	Observed streamflow	$mm\ day^{-1}$
q_{obs}	Flow percentile of observed streamflow	$mm\ day^{-1}$
Q_{sim}	Simulated runoff	$mm\ day^{-1}$
q_{sim}	Flow percentile of simulated runoff	$mm\ day^{-1}$
Q_{xx}/q	Flow percentile	$mm\ day^{-1}$
R^2	Coefficient of determination	$-$
$SAND$	Soil sand content	%
$SILT$	Soil silt content	%
SLO	Surface slope	$^{\circ}$
TA	Mean annual air temperature	K
URB	Fraction of urban area	$-$

Acknowledgements

First of all, I would like to thank my supervisor JProf. Dr. Andreas Hartmann for providing me the interesting topic and his outstanding and inspiring supervision. I am very grateful for Dr. Ingeborg de Graaf being my co-supervisor and for contributing her ideas to the topic, Dr. Michael Stölzle for sharing his advice on figure coloring, Dr. Hylke Beck from Princeton University for providing data, and the fellow students in the "blue container" for a pleasant working atmosphere and the mutual support during my part-time stay in the office. Not to be forgotten, I highly acknowledge Prof. Dr. Paolo Perona from the University of Edinburgh, who was not directly involved in this thesis, for his valueable scientific advice and being a great scientific inspiration. My parents are also thanked for supporting me throughout my studies. I would also like to thank Sinikka for proofreading and her critics on the thesis. Finally, a special thank to Jule for language proofreading and supporting me in general during the time of my thesis.

Abstract

The development of large-scale hydrological models (LHMs) has substantially enhanced the global prediction of water resources. Large-scale models may have, however, a strongly limited prediction performance. While traditional benchmarking efforts have given proof of the existence of model errors, to date, little empirical evidence exists on the direct quantitative link between the model errors and the surrounding climatic and physiographic settings. Hence, our aim was to identify settings in which model errors are embedded and examine the control mechanism of climatic and physiographic characteristics. To achieve this we systematically compared daily runoff simulations (1979-2012; 0.5°) of the ensemble mean from 10 state-of-the-art LHMs, all driven by the WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset, with 3653 observed streamflow time series (2-100 000 km^2). We, then, combined a clustering approach with a regression analysis and a random forest approach showing that model errors were linked to variables describing snow, evapotranspiration, soil and geologic characteristics. We found that errors originate from inadequacies of the corresponding model routines. Besides that, we ascertained "climatic" control of WFDEI P data on errors which we separated from structural errors by employing the clustering. Overall, the presented analysis proved to be a useful tool for advancing model development by identifying deficient model structures. The study emphasises the importance of the statistical approach allowing for concise insights into error control parallel to traditional benchmarking efforts.

Keywords: Large-scale models, Model evaluation, Global, Machine learning, Regression analysis

Zusammenfassung

Die Entwicklung großskaliger hydrologischer Modelle (LHMs) hat die globale Vorhersage von Wasserressourcen erheblich verbessert. Allerdings können großskalige Modelle eine stark eingeschränkte Vorhersageleistung haben. Während klassische Benchmarking-Bemühungen die Existenz von Modellfehlern belegen, gibt es bisher wenig empirische Belege für den direkten quantitativen Zusammenhang zwischen Modellfehlern und den umgebenden klimatischen und physiographischen Rahmedingungen. Unser Ziel war es daher, Rahmedingungen zu identifizieren, in die Modellfehler eingebettet sind, und den Kontrollmechanismus der klimatischen und physiographischen Eigenschaften zu untersuchen. Dafür verglichen wir systematisch die täglichen Abflusssimulationen (1979-2012; 0.5°) des Ensemblemittelwertes von 10 state-of-the-art LHMs, die alle durch den meteorologischen Datensatz WATCH Forcing Data ERA-Interim (WFDEI) angetrieben wurden, mit 3653 beobachteten Abflußzeitreihen (2-100 000 km^2). Anschließend kombinierten wir einen Clustering-Ansatz mit einer Regressionsanalyse und einem Random Forest Ansatz. Hier zeigen wir, dass Modellfehler mit Variablen zur Beschreibung von Schnee, Evapotranspiration, Boden und geologischen Eigenschaften verknüpft sind. Wir haben festgestellt, dass Unzulänglichkeiten in den entsprechenden Modellroutinen dafür verantwortlich sind. Außerdem stellten wir eine "klimatische" Kontrolle der WFDEI P-Daten auf die Fehler fest, die wir durch den Einsatz des Clusterings von strukturellen Fehlern getrennt haben. Insgesamt erwies sich die vorgestellte Analyse als nützliches Werkzeug für die Weiterentwicklung der Modelle durch die Identifizierung mangelhafter Modellstrukturen. Dabei steht im Vordergrund, dass der statistische Ansatz prägnante Einblicke in die Kontrolle der Fehler parallel zum klassischen Benchmarking ermöglicht.

Stichworte: Großskalige Modelle, Modellevaluierung, Global, Machine Learning, Regressionsanalyse

1 Introduction

In the late 1980s and early 1990s the first detailed global water resources assessments were carried out stating a shortage of global water resources (Bierkens 2015). These early exertions considered only statistics of water use (e.g., AQUASTAT) and observations of meteorological and hydrological variables. Thus, with such heterogeneous information the assessment remained very rough as the statistics provided mostly national averaged values (e.g., water use). Furthermore, the quality and density on which the statistics built up might have differed strongly between countries. Unfortunately, these issues are still present today.

The emerging awareness of the shortage of global water resources and the need for an improved assessment, however, lead shortly thereafter to the development of the first large-scale hydrological models (LHMs). In contrast to the first assessment efforts embedding LHMs, water resources could then be assessed more homogeneously. Among those, WaterGap (Alcamo et al. 1997), WBM (Vörösmarty et al. 1998) and MacPDM (Arnell 1999) are considered important pioneers in the field of global hydrological modeling. The very basic idea of these models was to determine the water availability globally. This was done by accumulating runoff over a stream network. In addition to that, WaterGap and MacPDM implemented first routines to calculate the water demand. While subtracting the demand from the available water, the water stress can be estimated. Since then, various LHMs have appeared and have undergone several rounds of improvement, increasing both functionality and resolution of the models. For a detailed genealogy of these models and their functions we refer to Bierkens (2015).

Apart from global water resources assessment LHMs have been applied for many purposes including, but not restricted to, flood and drought hazard assessment (Hirabayashi et al. 2013; Pappenberger et al. 2012; Tallaksen and Stahl 2014; Ward et al. 2013), global groundwater depletion (Gleeson et al. 2012; Wada et al. 2010) and assessing hydrological impacts of climate change (Pokhrel et al. 2013; van Vliet et al. 2016; Wada et al. 2012). In this respect, LHMs are often integrated as a supportive tool into decision making (e.g., Yu et al. 2015). Yet, since in general all models can only reproduce the real-world imperfectly, they may have a strongly limited predic-

tion performance. Thus, it is essential that the limited prediction performance is not neglected by the decision making processes. However, despite the knowledge about this limited prediction performance systematic methods to benchmark those models have rarely been deployed until today. Originally, the first model evaluation efforts had been dedicated to catchment-scale models and an enormous amount has been published on that ever since (e.g., Beven 2011). Although model evaluation of LHMs is subject of an emerging field of research, the number of studies cannot compete yet with those of catchment-scale models. Reasons for that are computational challenges on the one hand, lack of large datasets to characterise climatic and physiographic settings (e.g., land use, soils, surface elevation, etc.) (Sperna Weiland et al. 2015) on the other hand. Furthermore, observed data to which simulations can be compared to are only available for certain catchments and their accuracy and reliability is subject to global variations (Sperna Weiland et al. 2015). As a consequence, evaluation is spatially restricted to those catchments for which observations are available (Sperna Weiland et al. 2015).

1.1 State of the Art

Benchmarking models using independent data sources is paramount for advancing model development, rejecting deficient model structures and quantifying model credibility (Beck et al. 2017a). In order to quantify these uncertainties several model intercomparison initiatives have been established (e.g., WaterMIP: Haddeland et al. 2011; ISI-MIP: Schellnhuber et al. 2014). These initiatives have yielded numerous multi-model evaluation focusing on hydrological variables. Among those, runoff is one of the most useful variables for evaluation since it reflects the integrated catchment response and thereby the involved hydrological processes (Beck et al. 2017a). Moreover, observed runoff data is readily available for many catchments by public data bases. At present, 22 large-scale studies evaluating the runoff simulations of multiple models exist (Table 1.1). These studies typically consider LHMs. Within the realm of LHMs, it is necessary to distinguish between two classes of models: global hydrological models (GHMs) and land surface models (LSMs), the latter ones being extended with hydrological schemes. In contrast to GHMs, LSMs are not deliberately dedicated to the estimation of daily runoff. Originally, they have been developed to reproduce soil-atmosphere interactions (Beck et al. 2017a). As a consequence, they are often coupled with a separate routing scheme to convert simulated runoff to streamflow. Whether studies are using such a routing model is

marked explicitely in Table 1.1.

In the following, we briefly describe the studies listed in Table 1.1 with respect to their methodology and their main findings and conclusions. It must be noted that the individual study setup will not be mentioned explicitely. Instead we refer to Table 1.1. From the studies listed in Table 1.1 9 covered the continental scale and 13 covered the global scale. Concerning the continental scale only studies for the European and North-American continent exist, the latter ones (3 in total) are focusing on the conterminous USA. In this respect, Lohmann et al. (2004) were the first who conducted a large-scale study in which they evaluated simulated runoff of medium-sized catchments from four LSMs. For this purpose, they used the relative bias of the average runoff and the Nash-Sutcliffe-Efficiency (NSE) (Nash and Sutcliffe 1970). With their findings they concluded a runoff underestimation of all models in areas with significant snowfall. Xia et al. (2012) conducted their study with the exact same set of models as Lohmann et al. (2004), but in contrast they used a different forcing and they added the anomaly correlation as a further evaluation metric. Their findings reveal a good model performance in the eastern USA and the west coast of the USA. Having a closer look, a poor NSE was found for predicting daily streamflow, but they could prove the presence of an anomaly correlation. Using the ensemble mean they improved the predictive accuracy which outperformed the single models.

The most recent study of Melsen et al. (2018) investigates the uncertainty of hydrologic projections. In their analysis they incorporated the sign of change of the average annual runoff and the discharge timing between two time periods, 1985-2008 and 2070-2100. This could reveal model uncertainties in regions where snow processes and aridity are dominant. As a major source of uncertainty they identified the forcing by the Global Circulation Models (GCMs). Similarly, Milly et al. (2005) examined global pattern of trends in streamflow of large-sized basins and projected these trends. This was also the first noteworthy study at a global-scale comparing runoff estimates to observations; their modeling approach, however, is rather less "hydrologic" and not in line with the studies presented in Table 1.1. This is the case because runoff is being simulated by routing precipitation of several climate models by a simple linear reservoir. They used, however, the ensemble mean of models which was generally capable in reproducing the observed trend patterns in runoff. Consistent disagreement in sign of the trend was found in Central America and northern South America, northeastern Europe, and central and southeast Asia.

Five continental studies were carried out for Europe. Compared to the other two

Table 1.1: Overview of all currently available large-scale (continental to global) studies evaluating simulated runoff of multiple models, sorted by region and publication date. The table is adapted from Beck et al. (2017a).

Study	Region	Number of (identical) models	Number of catchments (size range)	Time period	Model resolution	Evaluation timescale(s)	Forcing	Routed runoff
Lohmann et al. (2004)	Cont. USA	4 (0)	1145 (23 - 10 000 km^2)	1996 - 1999	1/8°	Daily, monthly, annual, long-term	EDAS[a]	yes
Xia et al. (2012)	Cont. USA	4 (0)	969 (23 - 1 353 280 km^2)	1979 - 2007	1/8°	Daily, weekly, monthly, annual, long-term	NLDAS[b]	no
Melsen et al. (2018)	Cont. USA	3 (0)	605 (4 - 25 800 km^2)	1985 - 2008, 2070 - 2100	0.5°	Long-term	GCMs	no
Prudhomme et al. (2011)	Europe	3 (2)	579 (< 1000 km^2)	1963 - 2001	0.5°	Daily	WFD[c]	no
Gudmundsson et al. (2012a)	Europe	9 (4)	426 (< 4000 km^2)	1963 - 2000	0.5°	Daily, annual, long-term	WFD[c]	no
Gudmundsson et al. (2012b)	Europe	9 (4)	426 (< 4000 km^2)	1963 - 2000	0.5°	Annual, long-term	WFD[c]	no
Greuell et al. (2015)	Europe	5 (1)	46 (9948 - 658 340 km^2)	1970 - 2000	0.5°	Daily, monthly, annual, long-term	E-OBS, WFDEI[d], WFD[c]	no
Gudmundsson and Seneviratne (2015)	Europe	9 (4)	426 (< 4000 km^2)	1963 - 2000	0.5°	Monthly, annual, long-term	WFD[c]	no
Milly et al. (2005)	Global	12 (0)	165 (> 50 000 km^2)	1900 - 1998, 2041 - 2060		Long-term	GCMs	yes
Decharme and Douville (2006)	Global	6 (0)	80 (100 000 - 4 758 000 km^2)	1982 - 1995	1°	Daily, monthly	GSWP-2[e]	yes
Decharme and Douville (2007)	Global	6 (0)	80 (100 000 - 4 758 000 km^2)	1982 - 1995	1°	Monthly	GSWP-2[e]	yes
Decharme (2007)	Global	2 (0)	80 (100 000 - 4 758 000 km^2)	1982 - 1995	1°	Monthly	GSWP-2[e]	yes
Materia et al. (2010)	Global	13 (1)	30 (82 000 - 4 677 000 km^2)	1986 - 1995	1°	Monthly	NCEP-DOE	yes
Zaitchik et al. (2010)	Global	4 (0)	66 (19 000 - 4 600 000 km^2)	1979 - 2007	1°	Daily, monthly, annual	GLDAS[f], Princeton[g]	yes
Haddeland et al. (2011)	Global	11 (3)	8 (650 000 - 4 600 000 km^2)	1985 - 1999	0.5°	Monthly, annual	WFD[c]	no
Zhou et al. (2012)	Global	14 (0)	150 (not specified; $\gg$10000 km^2)	1986 - 1995	1°	Annual	GSWP-2[e]	no
van Dijk et al. (2013)	Global	5 (1)	6192 (10 - 10 000 km^2)	1979 - 2008	1°	Monthly	Princeton[g], GLDAS[f]	no
Beck et al. (2015)	Global	4 (3)	4079 (10 - 10 000 km^2)	1979 - 2015	0.25 - 1°	Daily, long-term	ERA-Interim[g], GLDAS[f], Princeton[g]	no
Yang et al. (2015)	Global	7 (1)	16 (135 757 - 3 475 000 km^2)	1981 - 2010	0.5°	Monthly, annual, long-term	CRUNCEP	yes
Zhang et al. (2016)	Global	4 (0)	644 ($\gg$ 2000 km^2)	1981 - 2010	0.5°	Monthly, annual	Princeton[g]	no
Beck et al. (2016)	Global	10 (10)	1113 (10 - 10 000 km^2)	1979 - 2012	0.5°	Daily, 5-day, monthly, long-term	WFDEI[d]	no
Beck et al. (2017a)	Global	10 (10)	966 (1000 - 5000 km^2)	1979 - 2012	0.5°	Daily, 5-day, monthly, annual, long-term	WFDEI[d]	no
This study	Global	10	5482 (2 - 100 000 km^2)	1979 - 2012	0.5°	Daily, long-term	WFDEI[d]	no

[a]Rogers et al. (1999) [b]Cosgrove et al. (2003) [c]Weedon et al. (2011) [d]Weedon et al. (2014) [e]Dirmeyer et al. (2006) [f]Rodell et al. (2004) [g]Sheffield et al. (2006) [g]Dee et al. (2011)

regions, those studies all share the same model resolution and their forcings are almost the same (see Table 1.1). Prudhomme et al. (2011) posed the question on how well large-scale models reproduce regional hydrological extremes in Europe. They applied the regional deficiency index and the regional flood index representing low flows and high flows, respectively. The indexes derived from observed and simulated time series are evaluated by the relative mean error, the ratio of the standard deviation of simulated and observed runoff and the Spearman correlation. As a result, they recognized a skill in the models for reproducing the spatiotemporal evolution of hydrological extremes. Regarding the low flows the models capture the broad-scale characteristics, whereas deficiencies are found for high flows related to the spatial resolution in the forcing. Instead of focusing solely on hydrological extremes, Gudmundsson et al. (2012b) covered the entire flow range of the hydrograph with their study by comparing the model simulations to five observed runoff percentiles. They carefully analysed spatially aggregated annual time series of the five flow percentiles. On the one hand, the models are able to seize the interannual variability, which was reflected in the Spearman correlation, on the other hand, the relative bias in the mean and the standard deviation is diverging between models emphasising model uncertainties. In particular, the greatest divergence is encountered for low flows. The ensemble mean exhibited a good performance. Another study conducted by Gudmundsson et al. (2012a) evaluated LHMs with respect to the seasonal runoff climatology. For their setup they utilized the same models, catchments, investigated time period and evaluation metrics as in Gudmundsson et al. (2012b) (see Table 1.1). In addition to that, they were the first to described the influence of catchment characteristics and climatic conditions on model errors quantitavely. For that, the Spearman correlation was computed between the evaluation metrics and the catchment characteristics (e.g., mean catchment elevation) and climatic conditions (e.g., observed mean annual temperature). When analysing this they obtained a positive correlation between the difference in grid cell elevation and catchment elevation and the relative bias in the mean and a strong negative correlation between the relative bias in the mean and the runoff ratio. Their findings illustrate that models perform poorly for snow-influenced regions, while they perform well in all other regions. Although large differences are present among the models, the overall good performance of the ensemble mean can be emphasized.

Using the same study setup (see Table 1.1) like the two previously described studies Gudmundsson and Seneviratne (2015) applied a machine learning approach. In that respect, they employed random forests to estimate monthly runoff. The

statistical model included climate (e.g., temperature) and land parameters (e.g., slope) for the estimation. The results indicate that the model is capable of reproducing monthly runoff estimates with reasonable accuracy. Although the reasonable skill in runoff estimation was stated, uncertainties in model input data are emphasized by the authors, which can have a remarkable impact on the model outcome. Another machine learning approach is provided by Beck et al. (2015); they used artificial neural networks (ANNs) to generate global maps of streamflow characteristics based on climate and physiographic characteristics. A comparison of these maps with the simulations of 4 LHMs unveiled a weak performance of the LHMs in simulating baseflow recession rate, an early bias in discharge timing and an underestimation of runoff over mountain ranges. Additionally, they conducted a simple linear regression between streamflow characteristics and climate and physiographic characteristics. According to their findings, climate and topographic predictors are more important than the ones related to soil and geology.

At a global scale, Decharme and Douville (2006), Decharme (2007) and Decharme and Douville (2007) can be mentioned as the first studies evaluating mature LHMs. The setup of their studies, in which they focused on different issues, was identical. Decharme and Douville (2006) aimed to unravel the uncertainties in the forcing and their impacts on the hydrologic simulations. Taking the ratio of mean simulated and observed discharge and the NSE as evaluation metrics they could provide evidence for systematic errors in simulated discharge over the mid and high latitudes caused by overestimation of precipitation in the forcing. Analogously, Materia et al. (2010) subjected the sensitivity of simulated river discharge similarly to Decharme and Douville (2006), but they additionally appended land surface representations into their considerations. Although they used a different meteorological forcing data, their findings are in agreement insofar as simulated discharge is most sensitive to variations in precipitation.

Decharme (2007) further validated simulated runoff against monthly observed values. In comparison to Decharme and Douville (2006) they added the square correlation between simulated and observed monthly anomalies to the evaluation procedure. A simulated late snow melt uncovered the inadequate representations of snow melt processes in the selected models. Further limitations are given by the river routing scheme used to convert simulated runoff to streamflow in which, for example, seasonal floodplains are not considered. Zaitchik et al. (2010) evaluated simulations of four LSMs extensively by embedding mean annual discharge, seasonal and intraseasonal variability, interannual variability and the timing of peak

flow. Their evaluation highlights diverging accuracy of streamflow estimates related to geographic patterns. In particular, the models have the tendency to underestimate discharge in tropical regions. Mostly, this is caused by an underestimation of the precipitation in the forcing. By contrast, high-latitudes exhibit a poor timing in peak flow caused by an early onset of the snow melt where additionally simulations underestimate river discharge. Furthermore, Zaitchik et al. (2010) could show that the model accuracy is bound to the choice of the atmospheric dataset. The study of Haddeland et al. (2011) elaborates on the multimodel (dis)agreement in estimating the global terrestrial water balance of eight very large basins. For that cause, they compared mean annual and mean monthly values of water balance variables. No major differences in interannual variation of runoff could be found between models despite different runoff schemes. However, a high disagreement for runoff was observed in tropics and arid areas where in the latter case runoff is overestimated. They further stress a high inter-model disagreement for snow-influenced catchments due to different model implementations of snow hydrology. Zhou et al. (2012) benchmarked mean annual simulated runoff of 14 LSMs. In their approach they made use of the relative bias of the mean annual runoff, the coefficient of determination and the NSE. Large positive biases prevailed in northern high-latitudes originating from an overestimation of the precipitation in the forcing. Conversely, large negative biases for the Amazon and Orinoco region were found due to underestimated precipitation. They further provide evidence of large biases in regions with low mean annual precipitation. In particular, wet basins with a large baseflow ratio exhibit smaller biases than wet basins with small baseflow ratio. This error may be assigned rather to the model structure than to the forcing. Yang et al. (2015) assessed discharge simulation in dynamic global vegetation models (DGVMs). They could give proof for a good reproduction of the seasonal runoff cycle by the models at low- and mid-latitudes. Peak discharge in high latitudes was underestimated. Generally, mean annual runoff was understimated while interannual variability was well simulated. The results of Zhang et al. (2016) confirmed these results since the models are capable of reproducing seasonal and interannual variability. Moreover, simulations correlate reasonably with observations. Nevertheless, models perform poorly in simulating monthly and annual runoff. Recently, the study of Beck et al. (2017a) evaluated the runoff of 10 state-of-the-art LHMs. They focused on medium-sized catchments which used a broad range of performance metrics related to important aspects of the hydrograph. They revealed an early bias in spring snow melt peak due to an underestimation of precipitation in the forcing and misrepresentation of

certain processes in snow hydrology. Further precipitation biases in the forcing are present in (semi-)arid regions and propagates into the simulated runoff.

The two studies of van Dijk et al. (2013) and Beck et al. (2016) are not primarily designated to the evaluation of simulated runoff. Instead, LHM runoff estimates are used to assess and validate their novel developed approaches. van Dijk et al. (2013) implemented a seasonal streamflow forecasting system. In terms of predictive accuracy, compared to four LMSs, the system performed equally well or slightly better. The simulated runoff used in the previously described studies was generated by uncalibrated LHMs. Some of the models in Beck et al. (2017a) are called "calibrated". This notion should be considered with caution. Calibration of LHMs is not comparable to the calibration procedure applied to catchment-scale models. Different from that is the globally calibrated HBV model by Beck et al. (2016). They presented a global-scale regionalisation scheme for the first time. Using an aggregate objective function consisting of runoff signatures and goodness-of-fit measures they could illustrate that HBV with regionalised parameters outperformed nine state-of-the-art LHMs.

Summing these studies up we could identify some common deficiencies in simulating runoff with LHMs. Several studies attested the LHMs a poor performance in snow influenced regions (Beck et al. 2017a; Decharme 2007; Decharme and Douville 2006; Gudmundsson et al. 2012a; Haddeland et al. 2011; Lohmann et al. 2004; Melsen et al. 2018; Zaitchik et al. 2010). This was caused either by underestimation of precipitation in the forcing (e.g., Beck et al. 2017a; Gudmundsson et al. 2012b) or shortcomings of the models in representing snow dynamics (e.g., Zaitchik et al. 2010,Gudmundsson2012a). Further, poor performances were found for (semi-)arid regions (Beck et al. 2017a; Haddeland et al. 2011; Melsen et al. 2018) and tropical regions (Haddeland et al. 2011; Zaitchik et al. 2010) where wrong precipitation in the forcing translated into model errors. Furthermore, inadequate model structures (e.g., storage routine) were found to be responsible for underestimation and overestimation (e.g., Gudmundsson et al. 2012b; Zhou et al. 2012), respectively. A second common conclusion of the studies is that the ensemble mean can lead to an improvement of the predictive accuracy. Either it performs slightly worse than the best model (Beck et al. 2017a; Gudmundsson et al. 2012b) or it even outperforms the best model (Xia et al. 2012). However, many of these studies used either a relatively small amount of time series (≤ 200) or they only evaluated monthly or annual mean runoff (Beck et al. 2017a). Moreover, many of them incorporated only a few LHMs (≤ 5) or evaluation metrics (≤ 2) (Beck et al. 2017a) where the relative

bias of the average simulated and observed runoff, the Pearson correlation and the NSE are altogether frequently used.

1.2 Research Objectives

All these studies somehow addressed the model errors, but to our knowledge none of the studies has profoundly answered the question how or by what the model error is controlled quantitavely. To date, only the study of Gudmundsson et al. (2012a) has already put some effort in this direction. They linked climate and physiographic characteristics to streamflow charcteristics and model errors in a bivariate way. Yet, eventually, a bivariate approach might not account for the complex interplay of climatic and physiographic characteristics. A multivariate approach might be more appropriate for allowing such an interplay. Beck et al. (2013) and Beck et al. (2015) used similar climatic and physiographic catchment descriptors to investigate their control mechanisms on streamflow characteristics and baseflow characteristics, respectively. For this purpose, both studies implemented a bivariate and multivariate statistical approach.

Thus, the main research objective of this study is to reveal the linkage between model errors and the climatic and physiographic catchment descriptors with a bivariate regression analysis and random forest analysis. In order to identify the settings describing a hydrological system (Winter 2001) for which large scale models behave wrongly/incorrectly, we will divide the data according to the concept of hydrologic landscapes into subsets and run the analysis on each subset separately. The subsets might also facilitate the identification of potentially important hydrological processes. Distinguishing those subsets and the corresponding settings that cause the error, we hope to provide lead for modelers towards improvements.

In order to address this question, we systematically compare a global ensemble of hydrological simulations with a large dataset of observed streamflow time series (~5500) around the globe, mainly on a daily time scale. In a second step, we compare the model error to climatic and physiographic catchment descriptors. Our study is organised as follows:

- First, the runoff data both simulated runoff and observed streamflow are introduced.
- Secondly, we present the climatic and physiographic characteristics describing a hydrological system

- In the section of hydrologic landscapes the subsets are defined.
- We, then, introduce the metrics used for the model evaluation.
- Subsequently, the approach of statistical analysis is explained.
- Finally, we present and discuss the results.

To facilitate future benchmarking and to also enhance its comparability we provide parts from the methodology in the form of the Python-package *GHMeval*. This aims to encourage further model evaluation efforts.

2 Data

2.1 Forcing

Running GHMs successfully cannot be accomplished without the presence of a driver comparable to that of a car. Yet, unlike in a car this driver is not a person, but in the hydrological modelling jargon this driver is commonly known as forcing. In fact, the notion forcing typically refers to a meteorological dataset which comprise exogenous factors like precipitation or temperature which force the hydrologic system. This dataset is usually a reanalysis product. The term reanalysis describes a data assimilation scheme on heterogeneous distributed data. The reanalysis is essential since the structure of LHMs asks for spatio-temporally distributed input data and observations of meteorological variables are unevenly distributed over time and space. Currently, several global forcing datasets (e.g., Dee et al. 2011; Dirmeyer et al. 2006; Sheffield et al. 2006; Weedon et al. 2011, etc) exist. Each comes along with its own caveats (e.g., Rust et al. 2015). In our study the LHMs were all consistently forced by the daily 0.5° WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset (Weedon et al. 2014) that covered the period between and including 1979 and 2012. The dataset contains both 3-hourly time intervals and daily time intervals. WFDEI has been generated on the basis of the ECMWF ERA-Interim reanalysis (Dee et al. 2011) and is corrected with the CRU dataset (Harris et al. 2014). By doing that, the effects of incorrect elevation and monthly bias have been reduced. Although WFDEI scheme leads to an improved forcing, one needs to keep in mind that there exist a number of problems when it comes to application. In this respect seven issues are identified according to Schellekens et al. (2017) and Beck et al. (2017a):

1. Unrealistic high rainfall in Gabun, Africa likely due to a unit error in the reported precipitation
2. Concerns about the energy forcing terms of WFDEI over the Amazon region where the average incident longwave radiation is understimated and the average incident shortwave radiation is overstimated. The spurious estimation is attributed to the ERA-Interim dataset.

3. Incoming radiation noise at night time (about 0.05 $W\ m^{-2}$) for a certain number of time steps. This originate also from ERA-Interim dataset.
4. Large positive values of the average incident shortwave radiation ($> 5\ W\ m^{-2}$) at night time
5. Substantial conversion of liquid precipitation (in ERA-Interim) into snowfall (in WFDEI) for nine grid cells
6. P underestimation in the Rocky Mountains due to missing correction for orographic effects
7. P overestimation in the northern Great Plains

These issues are due to their influence on the model outputs very important as far as the interpretation of the results is concerned.

2.2 Simulated Runoff

In hydrological modeling the simulated runoff is one of the main model outputs because of its reflection of the integrated catchment response (see Chapter 1). Instead of taking the simulated runoff of a single model we here use an ensemble of models. Ensembles are widely used in earth system sciences (Beck et al. 2017a; Fowler and Ekström 2009; Guilyardi 2006; Krishnamurti et al. 2000; Schellekens et al. 2017). Typically, they contain the outputs from different models or from a single model with different realizations (Beck et al. 2017a). One advantage of this technique is that an improvement of the model predictive accuracy can be achieved even though less accurate models are included in the ensemble (Gudmundsson et al. 2012b; Milly et al. 2005; Xia et al. 2012). Beck et al. (2017a) proved the suitability of the ensemble for runoff prediction although it performed slightly worse than the best model. For our study the ensemble comprises 10 state-of-the-art LHMs. From the ensemble we derived the ensemble mean runoff, which we call from here on simulated runoff, by averaging the output of the 10 models. The ensemble mean runoff is readily available from the eartH2Observe Tier-1 dataset (Schellekens et al. 2017), in which the simulated runoff ($kg\ m^2\ s^{-1}$) is provided unrouted and on a daily time scale. The models were run globally at a daily time step for the period 1979-2012 using the same forcing dataset described in chapter 2.1. The ensemble includes two classes of models the GHMs and LSMs (see Sect. 1.1). An overview of the models is given in Table 2.1. Note that some ensemble members, such as WaterGAP3, integrated

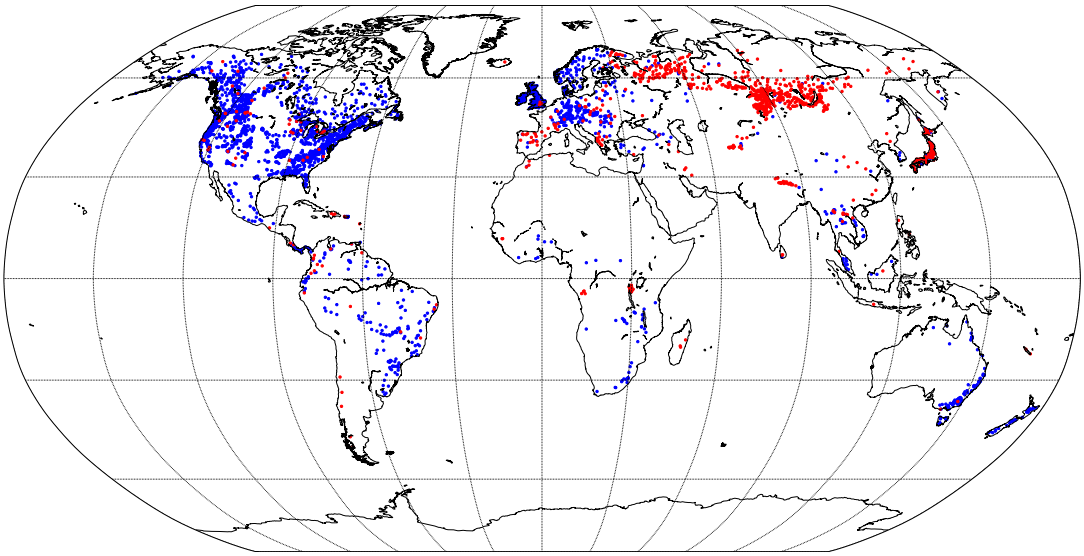


Figure 2.1: Location of the catchments and the corresponding time scales of the observed streamflow time series. The catchments are represented by its centroid. Catchments with daily time scale are indicated in blue and catchments with monthly time scale are in red.

anthropogenic water use. For details on the individual model spin-up procedure we refer to Schellekens et al. (2017). In addition to the mean ensemble runoff, we used the standard deviation ensemble runoff to describe the (dis)agreement in the ensemble (see Sect. 3.3).

Access to the data is given by the eartH2Observe Water Cycle Integrator (WCI: http://wci.earth2observe.eu) or by a THREDDS server (https://wci.earth2observe.eu/thredds/catalog/ens/wrr1/catalog.html) allowing direct download via OPeNDAP, WCS and HTTP (ftp is also supported). All files are in the Network Common Data Format (NetCDF).

2.3 Observed Streamflow

In order to evaluate the runoff estimates of the ensemble, we gathered observed streamflow time series and catchment boundaries from three different sources:

1. Global Runoff Date Centre (GRDC: http://ww.bafg.de/GRDC/)
2. Attributes of Gages for Evaluating Streamflow (GAGES)-II database (Falcone et al. 2010)
3. Australian streamflow data compilation (Peel et al. 2000)

The streamflow records are observed either at a daily or monthly time scale. At present, this dataset with approximately 5500 both near-natural and anthropogenically influenced catchments provides to our knowledge the greatest possible amount of observed streamflow time series one can publicly have access to. Since the performance metrics in Section 3.3 require rather daily values than monthly values, we transformed monthly time series into daily time series using linear interpolation. The locations of the 3653 catchments which were finally included in our study are shown in Fig. 2.1. For more details on how we chose the catchments we refer to Section 3.3. From those 3653 catchments 2754 are measured at a daily time scale and 881 at a monthly time scale. Additionally, information on the temporal coverage of the observed streamflow data for the period of 1979-2012 is given in Figure 2.2 and on the catchment area in Figure 2.3. Moreover, their overall distribution is illustrated in Figure 2.4. The record length of the observed streamflow time series varies from 5 to 33 years with median record length of 20 years, whereas the catchment area ranges from 2 to 100 000 km^2 with median size of 1047 km^2.

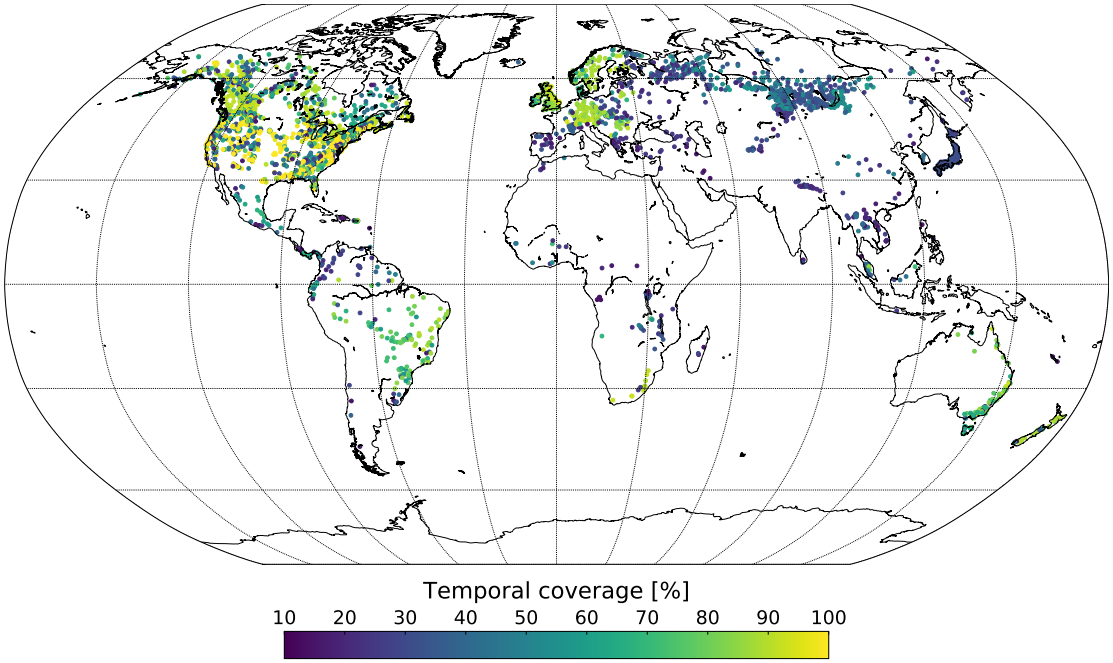


Figure 2.2: Global map of the temporal coverage (in %) of the observed streamflow time series between 1979 to 2012. Blue (yellow) displays low (full) temporal coverage.

Table 2.1: Overview of models and summary of processes included. The table is adapted from Schellekens et al. (2017).

Model	Model class	Interception	Evaporation	Snow	Soil layers	Groundwater	Runoff	Reservoirs/lakes	Routing	Water use	Time step
HTESSEL-CaMa	LSM	Single reservoir, potential evaporation	Penman-Monteith	Energy balance, 1 layer	4	No	Saturation excess	No	CaMa-Flood	No	1h
JULES	LSM	Single reservoir, potential evaporation	Penman-Monteith	Energy balance, 3 layers	4	No	Saturation and infilt. excess	No	No	No	1h
LISFLOOD	GHM	Single reservoir, potential evaporation	Penman-Monteith	Degree-day, 1 layer	2	Yes	Saturation and infilt. excess	Yes	Double kinematic wave linear cascade of reservoirs (sub-grid)	Yes	1 day
ORCHIDEE	LSM	Single reservoir structural resistance to evaporation	Bulk PET (Barella-Ortiz et al. 2013)	1 moisture layer, 1-5 thermodynamic layers	11	Yes	Green-Ampt infiltration	No	Travel time approach	irrigation only	900 s balance, routing energy 3 h
PCR-GLOBWB	GHM	Single layer, subject to open water evaporation	Hamon (tier 1) or imposed as forcing	Temperature based melt factor	2	Yes	Saturation excess	Tier 1 only lakes	TRIP with stream	Not in tier 1	1 day
SURFEX-TRIP	LSM	Single reservoir, potential evaporation	Penman-Monteith	Energy and mass balance, 12 layers	14	Yes	Saturation and infilt. excess	No	No	No	900 s for ISBA, 3600 s for TRIP
SWBM	GHM	No	Inferred from net radiation	Degree-day, 1 layer	1	No	Inferred from precipitation and soil moisture	No	No	No	1 day
W3RA	GHM	Gash event-based model	Penman-Monteith	Degree-day, 1 layer	3	Yes	Saturation and infiltration excess	No	Cascading linear reservoirs	No	1 day
WaterGAP3	GHM	Single reservoir	Priestley-Taylor	Degree-day, 1 layer	1	Yes	Beta function	Yes	Manning-Strickler	Yes	1 day
HBV-SIMREG	GHM	No	Penman 1948	Degree-day, 1 layer	1	No	Beta function	No	No	No	1 day

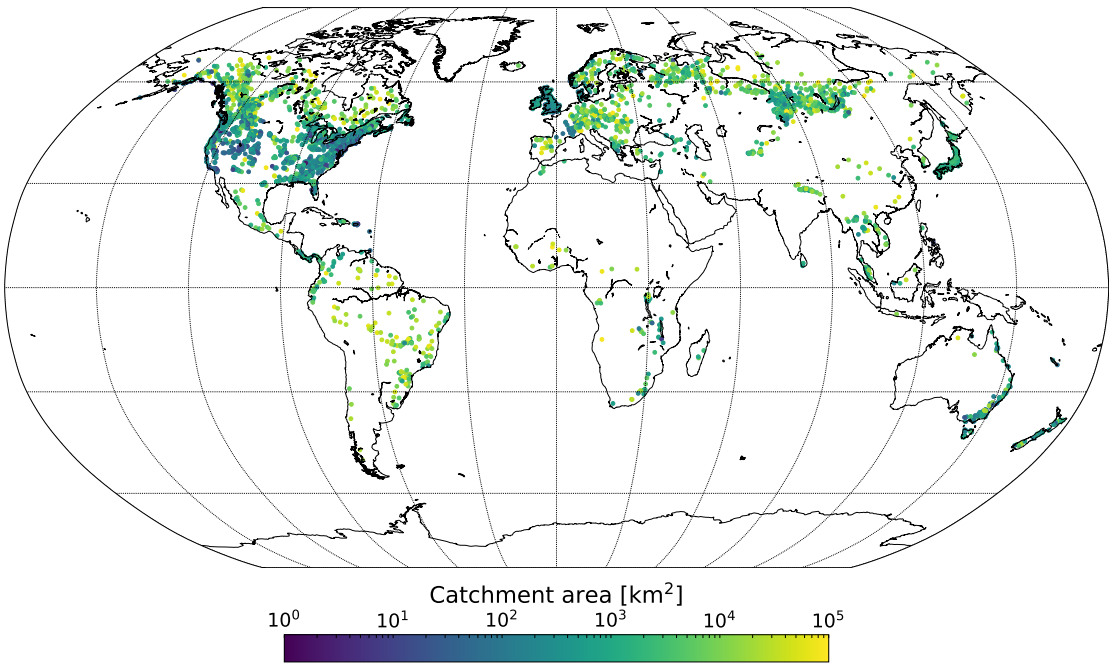


Figure 2.3: Global map of the catchment area (in km^2) of the observed streamflow time series. Blue (yellow) displays small-sized (large-sized) catchments.

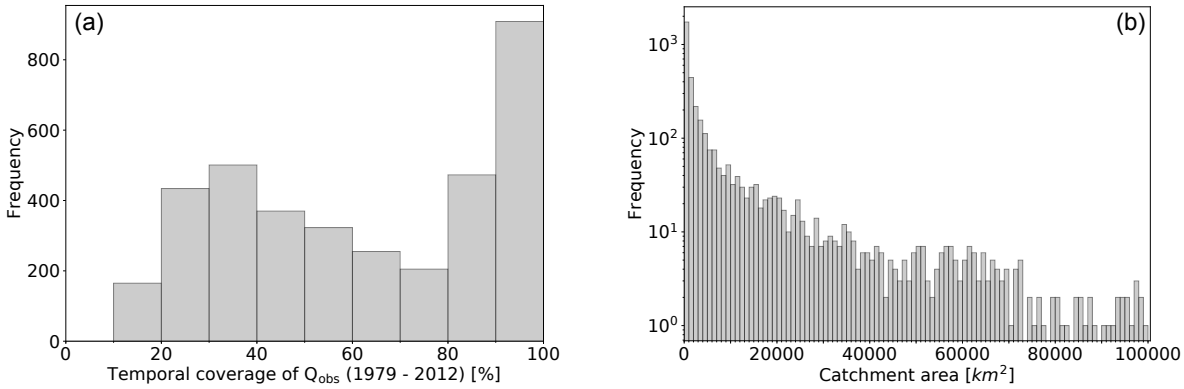


Figure 2.4: Distributions of temporal coverage (in %) of the observed streamflow time series between 1979 to 2012 (a) and the catchment area (in km^2) (b)

2.4 Climatic and Physiographic Characteristics

In Table 2.2 we present the climate and physiographic catchment characteristics. Although this selection is similar to the one in Beck et al. (2015), we, in contrast, modified and extended some characteristics, respectively. Precipitation was modified by using the most recent Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset (Beck et al. 2017b). In order to make the nonlinear AI a linear index we applied a logarithmic transformation. To exemplify this, distances would be nonlinear for an AI of 2 resulting from an PET twice as big as P while vice versa the AI

Table 2.2: Climatic and physiographic characteristics. The table is adapted from Beck et al. (2015).

Predictor	Unit	Description	Calculation and data source	Resolution
Climate				
AI	–	Aridity index	Calculated as $AI = PET/P$, where P is the mean annual precipitation and PET is the mean annual potential evaporation. We then applied a logarithmic transformation to AI. See P and PET for data sources.	~0.25°
P_{si}	–	Precipitation seasonality	Calculated following Walsh and Lawler (1981) as $P_{si} = P_{yr}^{-1} \sum \lvert P_m - P_{yr} \rvert / 12$, where P_{yr} and P_m are the mean annual and monthly precipitation, respectively, and the summation is over all months. See P for data source.	~0.25°
P	$mm\ yr^{-1}$	Mean annual precipitation	P is the mean annual precipitation derived from MSWEP (Beck et al. 2017b)	~0.25°
PET	$mm\ yr^{-1}$	Mean annual potential evaporation	Calculated from monthly values derived following the temperature-based approach of Hargreaves et al. (1985). See TA for data source.	~1 km
PET_{si}	–	Potential evaporation seasonality	Calculated following Walsh and Lawler (1981) as $PET_{si} = PET_{yr}^{-1} \sum \lvert PET_m - PET_{yr} \rvert / 12$, where PET_{yr} and PET_m are the mean annual and monthly potential evaporation, respectively, and the summation is over all months. See PET for data source.	~1km
$CORR$	–	Seasonal correlation between water supply and demand	Correlation coefficient calculated between monthly climate values of P and PET (Petersen et al. 2012). See P and PET for data sources.	~1 km
TA	K	Mean annual air temperature	WorldClim (Hijmans et al. 2005) and PRISM (Daly et al. 1994) for the United States.	~1km
PF	–	Permafrost abundance	National Snow and Ice Data Center vector map (Brown et al. 1997) with classes C, D, S, and I reclassified to permafrost abundances of 0.95, 0.70, 0.30, and 0.05, respectively.	~10 km
Topography (T)				
SLO	°	Surface slope	Consultative Group for International Agricultural Research (CGIAR) Consortium for Spatial Information (CSI) Shuttle Radar Topography Mission (SRTM), version 2.1 (Farr et al. 2007), for lat < 60°N, GTOPO30 (http://lta.cr.usgs.gov/GTOPO30) for lat > 60°N.	~90 m, ~1 km
$ELEV$	$m\ MSL$	Surface elevation	CSI SRTM, version 2.1 (Farr et al. 2007), for lat < 60°N, GTOPO30 for lat > 60°N.	~90 m, ~1 km
Land cover (LC)				
fW	–	Fraction covered by lakes and reservoirs	World Wildlife Fund (WWF) Global Lakes and Wetlands Database (GLWD) level 3 (Lehner and Döll 2004).	~1 km
$NDVI$	–	Normalized difference vegetation index	Systéme Pour l'Observation de la Terre (SPOT) Vegetation (VGT) S10 10-day max value composites (www.vgt.vito.be), mean of 2005.	~1 km
fS	–	Fraction of snow cover	MODIS Aqua snow cover daily level 3 global climate modeling grid product (MYD10C1), version 5 (Hall et al. 2006), mean of 2003-13.	0.05°
$GLAC$	–	Fraction covered by glaciers	Randolph Glacier Inventory (RGI), version 3.2, glacier outlines.	~30 - 60 m
Geology and soils				
$PERM$	$log_{10}\ m^2$	Permeability of geology	Global permeability map (Gleeson et al. 2011).	~1 km
fLi_{xx}	–	Fraction of lithologic class	Global Lithological Map database v1.1 (Hartmann and Moosdorf 2012) where the subscript xx denotes the corresponding lithologic class: ev, evaporites; ig, ice and glaciers; mt, metamorphic rocks; nd, no data; pa, acid plutonic rocks; pb, basic plutonic rocks; pi, intermediate plutonic rocks; py, pyroclastics; sc, carbonate sedimentary rocks; sm, mixed sedimentary rocks; ss, siliclastic sedimentary rocks; su, unconsolidated sediments; va, acid volcanic rocks; vb, basic volcanic rocks; vi, intermediate volcanic rocks; wb, water bodies	~2 km
$SAND$	%	Soil sand content	SoilGrids1km (Hengl et al. 2014) version April 2014, mean over all layers.	~1 km
$SILT$	%	Soil silt content	SoilGrids1km (Hengl et al. 2014) version April 2014, mean over all layers.	~1 km
$CLAY$	%	Soil clay content	SoilGrids1km (Hengl et al. 2014) version April 2014, mean over all layers.	~1 km
Water use (WU)				
URB	–	Fraction of urban area	"artificial areas" class of the map from GlobCover (Bontemps et al. 2011) version 2.3	~300 m
IRR	%	Percentage of irrigated area	Global Irrigated Area Map (Siebert et al. 2013)	0.08333°

would be 0.5 if AI is not transformed. Additionally, geologic characteristics have been extended by joining the proportional area of the lithology. For this purpose the global lithological map (GLiM) dataset of Hartmann and Moosdorf (2012) was utilized. Moreover, we included characteristics about water usage, which is described by the percentage of urban (Bontemps et al. 2011) and irrigated areas (Siebert et al. 2013), respectively. As a consequence, we have 33 descriptors in total delineating the climate and physiographic catchment characteristics. Among those descriptors, eight are related to climate, three to topography, five to land cover, two to geol-

ogy, three to soils, and two to water use. The fraction of lithological class Li_{xx} is subdivided in 16 lithologic classes. Most of the data have a resolution of $\leq$1km and values, for example, mean annual air temperature represent average catchment characteristics. It must be noted that NDVI ranges from 0 to 255 (Hylke Beck, personal communication) and not as it is supposed to be from 0 to 1. This is due to its data format. Thus, a value of 255 is equal to 1.

3 Methodology

3.1 Hydrologic Landscapes: Concept and Classification

Understanding the landscape embedded into a catchment is crucial when it comes to identification of the driving hydrological processes. However, the perception of a landscape in its entire complexity and describing the underlying interactions of its features may be an unsolvable task. From an hydrologic perspective focusing on three landscape dimensions which consist of topography, geology and climate may be sufficient to identify the most important attributed processes. By unifying those three dimensions into a conceptual hydrologic framework Winter (2001) developed the well-known concept of hydrologic landscapes. The core of this concept is built up by a fundamental hydrologic landscape unit (FHLU). It is illustrated by the diagram in Figure 3.1; different landscapes can be categorized using the three landscape features topography, geology and climate with respect to the attributed movement of water. The study of Wolock et al. (2004) exemplifies how one can turn the concept into practice. They applied the concept to the conterminous USA. For that they first ran a principal component analysis to data describing the landscape before they applied a cluster analysis to the outcome of the principal component analysis. With this approach they could identify 20 hydrologic landscape regions.

In an analogous manner to Wolock et al. (2004), we also make use of the concept of hydrologic landscape with the goal to partition our dataset into subsets according to their relevant hydrological processes. We argue that this is more hydrologic and more suitable for process-based model evaluation than borrowing the popular Köppen-Geiger climate classification system. In this sense, the very recent study of Knoben et al. (2018) disproved the suitability of the Köppen-Geiger classification for hydrological studies. A hydrologic landscape is described through three dimsensions by aridity index (AI), surface elevation ($ELEV$) and permeability ($PERM$) (see Table 2.2). We, then, applied a K-means clustering to it. With means of the K-means clustering we could divide our dataset into K distinct and non-overlapping clusters (James et al. 2017). A pivotal step is the determination of the appropriate number of clusters K. In order to specify K we utilized the so-called “elbow

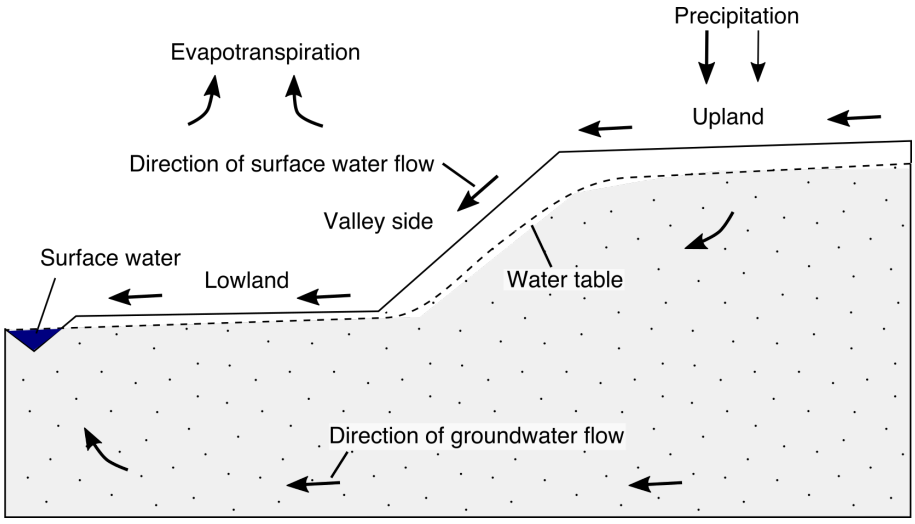


Figure 3.1: Diagram of a Fundamental Hydrologic Landscape Unit. The figure is adapted from Winter (2001).

method", a visual approach in which the sum of squared errors is plotted against the number of clusters. K is chosen when an distinctive elbow is perceptible, which means that at this point a greater K will reduce the sum of squared errors only marginally. Hartmann et al. (2015) came up with an smiliar approach defining typical karst landscapes. In our case each cluster represents a hydrologic landscape region (HLR).

3.2 Assignment of Simulated Runoff

Allowing a comparison of simulated runoff and observed streamflow, the simulated runoff in its gridded form has to be assigned to each gauged catchment. This is achieved similarly to Beck et al. (2017a). Depending on the match of grid cell centroid(s) and the catchment boundaries, we suggest here three options for assigning the simulated runoff:

1. Single grid cell centroid is enveloped by the catchment. The runoff of the single grid cell is assigned.
2. Multiple grid cell centroids are located within the catchment boundaries. The runoff of the multiple grid cells is first averaged before assigning it.
3. No grid cell centroid is within the catchment boundaries. The runoff is provided by the grid cell whose centroid is closest to the catchment's centroid.

Simulated runoff and observed streamflow are paired simultaneously with each other.

Thereby, both time series are equipped with the same length which is important for their comparability. Additionally, we assigned the ensemble standard deviation runoff to each catchment making use of the exact same procedure as described above. Note that simulated values and observed values are indicated in $mm\ d^{-1}$. The observed streamflow, originally indicated in m^3/s, is converted using the given catchment area.

3.3 Model Evaluation

In order to evaluate the model error we introduce some evaluation metrics here. Since we use an ensemble of 10 models, we also introduce an measure describing the inter-model disagreement. In particular, we calculate three evluation metrics:

1. Bias B_q (-) between simulated and observed values for 5 flow percentiles. High flows are characterised by 5 percentiles (Q_5), moderate high flows by 25 percentiles (Q_{25}), medium flows by 50 percentiles (Q_{50}), moderate low flows by 75 percentiles (Q_{75}) and low flows by 95 percentiles (Q_{95}). The definition of the flow percentiles corresponds to the statistical convention commonly used in Europe representing the exceedance frequencies (Gudmundsson et al. 2012b). We are calculating three sorts of biases, namely, a relative Bias $B_{rel,q}$ (-), a standardised Bias $B_{std,q}$ (-), and a standardised-square-rooted Bias $B_{std-sqrt,q}$ (-). The relative Bias is defined as

$$B_{rel,q} = \frac{q_{sim} - q_{obs}}{q_{obs}} \tag{3.1}$$

whereas $B_{std,q}$ and $B_{std-sqrt,q}$ can be expressed mathematically as

$$B_{std,q} = \frac{q_{sim} - q_{obs}}{\sigma_{obs}} \tag{3.2}$$

$$B_{std-sqrt,q} = \frac{\sqrt{q_{sim}} - \sqrt{q_{obs}}}{\sigma_{obs}} \tag{3.3}$$

where q represents the flow percentile and the sim and obs subscripts refer to simulated and observed runoff values, respectively. σ denotes the standard deviation of q_{obs} over all catchments and represents spatial variability across the landscape (Beck et al. 2017a). The B_q values range from $-\infty$ to $+\infty$, with lower values corresponding to better model performance. A negative

value suggests underestimation whereas a positive sign indicates the opposite. Since the majority of the catchments are small sized (see Figure 2.4b) we argue to use the relative bias in addition to standardised biases to give small catchments a greater weight. For those catchments the relative bias will be more pronounced.

2. Kling-Gupta-Efficiency (KGE). The KGE (Gupta et al. 2009) is a very popular efficiency measure widely used for model calibration. We, here, use a modified version introduced by Kling et al. (2012). By omitting the r term which represents the correlation coefficient between simulated and observed runoff we further modified the measure. We call this measure $KGE_{\gamma\beta}$ (-) and it is calculated as follows:

$$KGE_{\gamma\beta} = 1 - \sqrt{(\gamma - 1)^2 + (\beta - 1)^2} \tag{3.4}$$

where

$$\gamma = \frac{CV_{sim}}{CV_{obs}} = \frac{\frac{\sigma_{sim}}{\mu_{sim}}}{\frac{\sigma_{obs}}{\mu_{obs}}} \tag{3.5}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} \tag{3.6}$$

with σ as the standard deviation of the runoff and μ as the average runoff. The subscripts sim and obs denote simulated and observed runoff values, respectively. γ (-) is the variability ratio and β (-) represents the bias ratio. The maximum attainable value is 1. We used this modified version since it prevents a potential cross-correlation between the variability and bias ratio (Kling et al. 2012). As the model input comes in form of reanalysis data, we argue that the r term is biased because of discrepancies in real-world precipitation and precipitation in the reanalysis data. Also using the NSE would allow a intercomparibility to other studies because many of them used NSE for their evaluation (see chapter 1.1); yet, we omit this efficiency measure. This is because NSE is highly criticised (Schaefli and Gupta 2007) for being overly sensitive to the timing and the magnitude of peak flows. The metrics defined by equations (3.1), (3.3), and (3.4) refer hereafter to the model errors.

3. Coefficient of variation of simulated runoff at each flow percentile CV_q (-). Quantifying the inter-model (dis)agreement of the ensemble is achieved by

calculating the CV_q. It can be calculated because every time series of simulated runoff is paired simultaneously with the corresponding time series of standard deviation for the simulated runoff. Thus, at each flow percentile the related standard deviation can be extracted. We formulate CV_q as

$$CV_q = \frac{q_{sd}}{q_{sim}} \tag{3.7}$$

where the subscript *sd* denotes the standard deviation of the simulated runoff. A value of 0 reflects perfect inter-model agreement. The metric defined by equation (3.7) is hereafter called the inter-model disagreement.

We calculate the introduced evaluation metrics both for the entire dataset and each HLR. In order to select interesting HLRs we inspected the distributions of the model errors visually and chose those for which the error is likely not caused by the forcing.

For both the entire dataset and the selected HLRs we compared the distributions of the evaluation metrics in the form of boxplots. Additionally, the underlying median, mean, standard deviation, skewness and kurtosis of the distributions are computed. To examine whether the distribution of the evaluation metrics of the entire dataset and the HLRs are equal a two-sample Kolmogorov-Smirnov-Test (Conover 1971) was carried out. This is a two-sided test for the null hypothesis that two independent samples are drawn from the same continuous distribution. If the KS statistic is small or the p value is high, we cannot reject the hypothesis that the distributions of the two samples are the same.

Initially, we computed the evaluation metrics for the original dataset. Yet, we encountered some outliers in the B_{rel} which might originate from inaccuracies of streamflow measurements (Sperna Weiland et al. 2015). We excluded those catchments for which B_{rel} fall outside the range spanning from 1.5 times the inter-quantile range subtracted from the first quartile to 1.5 times the inter-quantile range added to the third quartile. By that around 1800 catchments had to be excluded. This radical step was necessary to prevent us from false conclusions on the one hand, and to make the results of our statistical analysis more reliable on the other hand.

3.4 Bivariate Regression: Climatic and Physiographic Controls on Model Errors

In order to scrutinize the climatic and physiographic control on the evaluation metric we conducted a bivariate regression analysis. Beck et al. (2013) investigated the same control mechanism. They, however, focused on catchments base flow characteristics. In a similar manner we therefore employed simple (non-)linear regression. Parameters of linear, exponential, logarithmic and power functions were fitted by least squares. The function reaching the highest coefficient of determination (R^2) was used to describe the relationship between climatic and physiographic characteristics and the evaluation metrics defined in the previous section. The R^2 was also used to quantify the strength of the relationship. 0 indicates that no relationship exists and 1 accounts for a perfect relationship. We define strength of the relationship as follows: 0–0.1 (no relationship), > 0.1–0.3 (weak), > 0.3–0.6 (moderate), > 0.6–0.9 (strong), and > 0.9 (very strong). In particular, the regression curve was not tested for significance (p value) since the p value may be misleading, especially when using large number of catchments (Beck et al. 2013). The bivariate regression analysis was carried out both for the entire dataset and the HLRs. Although the bivariate regression analysis is straightforward in terms of application and interpretation, one of the major disadvantages is that complex interplay of climatic and physiographic characteristics cannot be reflected.

In order to account for spatial scaling effects (Gudmundsson et al. 2012a), we ran a regression analysis between the evaluation metrics and catchment size for both the entire dataset and the HLRs.

3.5 Random Forest: Climatic and Physiographic Controls on Model Errors

Enabling such complex interplay we introduce a machine learning approach here. Machine learning is already widely applied in hydrology. Fields of application range from regionalisation approaches for the prediction in ungauged basins (e.g., Blöschl et al. 2013) to hydrologic impact studies (e.g., Bachmair et al. 2016) and classification approaches (e.g., Cloutier et al. 2008). Numerous machine learning methods exist; we, however, focus on random forest (RF) only, which has already been estab-

lished in the hydrologic community. For example, Bachmair et al. (2016) utilized RFs to predict drought impacts on the basis of drought indicators.

The RF technique, a supervised machine learning algorithm, was originally developed by Breiman (2001). In its core a RF consists of an ensemble of regression trees (James et al. 2017). For our analysis we applied the RF algorithm implemented in Python's machine learning library "Scikit-learn" (Pedregosa et al. 2011). In a random forest each regression tree is constructed on bootstrapped sub-sample of the dataset for which the pairing between predictors (climatic and physiographic characteristics) and predictand (evaluation metrics) is preserved. The term bootstrapping means that samples are drawn randomly with replacement from the dataset for which the sample size is about two-thirds of the sample size of the dataset (James et al. 2017). In contrast to the parametric bivariate regression analysis a non-paramteric regression tree applies a series of splitting rules to the bootstrapped sub-sample, i.e. each time a split in a tree is considered a sub-sample of m predictors is randomly chosen from the full set of p predictors (see Table 2.2) (James et al. 2017). Hence, splitting rules are defined such that they minimize the residual sum of squares. Thus, the prediction represents the average of observed values for the regions defined by the splitting rules. Finally, predictions are averaged over all trees.

We set the number of trees (n_estimators) used to build the random forest sufficiently large to 1000 to enable a good performance whereas the size of subsamples of predictors at each split (max_features) is set to $m \approx \sqrt{p}$. In this way the algorithm is forced to consider only small subsamples of predictors at each split. Typically, m is chosen small when we have a large number of correlated predcitors (James et al. 2017). As a consequence, the trees consider on average only $(p - m)/p$ predictors, which means that splitting rules will not solely be dominated by strong predictors (James et al. 2017). The remaining parameters take on the defaults. Using this parameterization RFs were, then, run for the entire dataset and each HLR.

To distinguish those climatic and physiographic characteristics which are best linked to the evaluation metric we used the unscaled and unconditional "permutation importance" measure described in Strobl et al. (2008). In particular, the measure quantifies the mean decrease in accuracy on the out-of-bag subsamples (i.e. samples which are not drawn by the bootstrap and consequently unseen by the RF) when a predictor is being perturbated. We, then, ranked the variable importance measure to identify important predictors. Moreover, to make the relationship between predictors and model output visible in a smiliar way like in Section 3.4 we computed the partial dependence of the predictions on important predictors (Hastie

et al. 2017). Partial dependence reflects the marginal effect of a given predictor towards the model outcome for which average model outcomes are derived for different values of the predictor (Hastie et al. 2017). In other words, the idea behind partial dependence is that it illustrates how the value of the predictor influences the model predictions after we have averaged out the influence of all other variables. For computational reasons partial dependence was only calculated for the four most important predictors: it was computed for an equally-spaced grid of size 30. The grid ranges from the minimum to the maximum of the given predictor. The partial dependence will be shown together with the deciles of the given predictor to include information on data density. Since the human perception is limited to low dimensions, we illustrate the single-variable partial dependence of the four highest ranked predictors in two dimensional line plots and the two-variable partial dependence for the two highest ranked predictors in form of a three dimensional surface plot. The latter one illustrates partial dependence of the prediction on joint values of the two most highest ranked predictors.

Note that due to the explorative character of our analysis we have not been conducting a cross-validation. Nonetheless, we show the R^2 computed on the out-of-bag subsamples to give evidence about the model accuracy (James et al. 2017). This gives an insight into the accuracy of the RF predictions. The computation of it is, in contrast to the cross-validation, very economic since the out-of-bag accuracy (R^2) is already included in the fitting procedure of the RF (Hastie et al. 2017). We define the model accuracy as follows: 0–0.1 (poor), $>$ 0.1–0.3 (fair), $>$ 0.3–0.6 (moderate), $>$ 0.6–0.9 (good), and $>$ 0.9 (excellent).

4 Results

4.1 Hydrologic Landscape Regions

Inspecting the elbow plot of the cluster analysis (Fig. 4.1) ten clusters can be determined. Yet, since this number of clusters does not result in plausible clusters (e.g., single clusters encompassed both humid and arid climates), we further increased the number of clusters to 12 resulting in coherent clusters for the three dimensions by which the clusters are described (Fig. 4.3) as well as by their locations (Fig. 4.2). On the basis of distinct cluster means (Table 4.2) and their corresponding standard deviations (Table 4.2) we defined 12 hydrologic landscape regions and attributed the assumed main hydrological processes with respect to their primary flow paths (Table 4.1). In terms of aridity index (AI) we distinguished between very humid ($AI < 0.5$), humid ($AI < 1$), sub-humid ($AI < 2$), semi-arid ($AI < 4$), and arid ($AI > 4$). In context of surface elevation ($ELEV$) we separated in plains ($ELEV <$ 1000), plateaus/low range mountains ($ELEV$ 1000 - 2000), and mountains ($ELEV$ $<$ 2000). The permeability of the geology ($PERM$) was classified into very permeable ($PERM <$ -12.3), permeable ($PERM <$ -13.5), impermeable ($PERM >$ -13.5) and very impermeable ($PERM >$ -14.8). Note that the term permeable

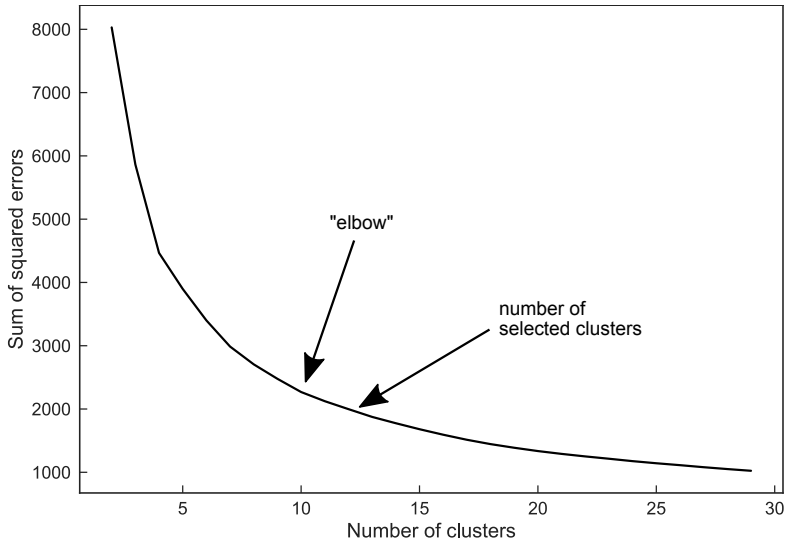


Figure 4.1: Elbow plot of sum of squared errors for K-means clustering

reflects higher permeability while impermeable indicates lower permeability. For the classification of *PERM* we followed Gleeson et al. (2011). The permeability of impermeable crystalline rocks, for example, are associated with -14.1 while highly permeable carbonate rocks are assigned a value of -11.8 (Gleeson et al. 2011). In case of the cluster mean being close to one of the thresholds and the cluster standard deviation not being small enough, we merged two classes (e.g., (sub-)humid or plains/plateaus).

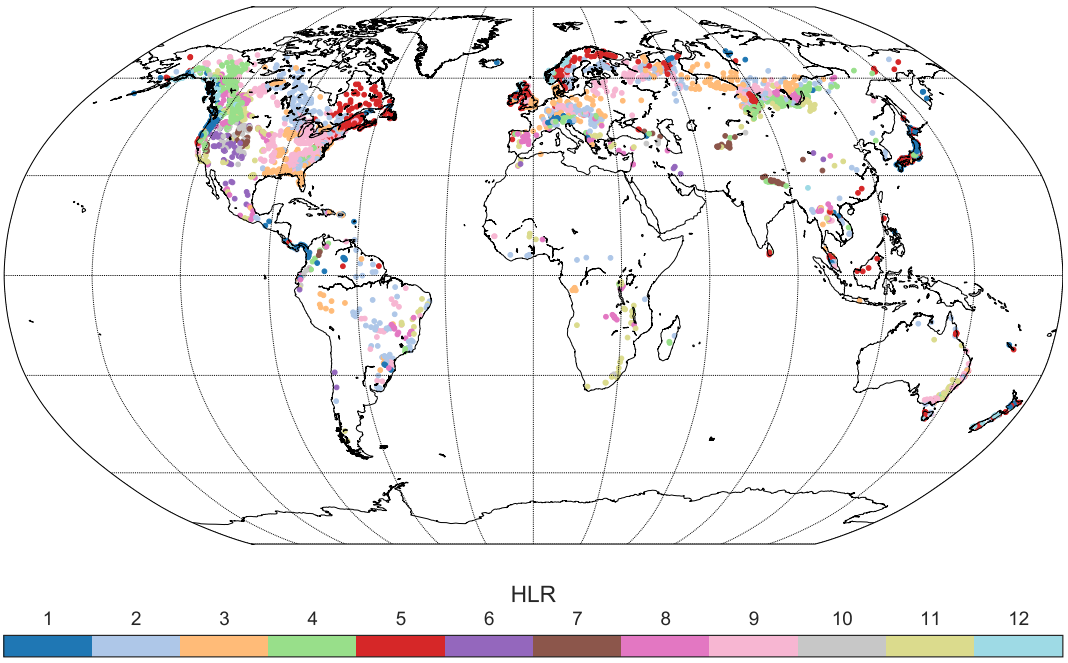


Figure 4.2: Global hydrologic landscape regions (HLR)

Table 4.1: Descriptions of hydrologic landscape region (HLR) and the assumed hydrologic flow paths

		Primary hydrologic flow paths		
HLR number	Description	Overland flow	Shallow ground water	Deep ground water
1	Humid plains with permeable bedrock		x	x
2	(Sub-)Humid plains with impermeable bedrock	x	x	
3	(Sub-)Humid plains with very permeable bedrock		x	x
4	Humid low range mountains with impermeable bedrock	x		
5	Humid plains with impermeable bedrock	x	x	
6	Subhumid and (semi-)arid mountains with permeable bedrock	x	x	
7	Humid mountains with impermeable bedrock	x		
8	Subhumid plains/plateaus with permeable bedrock		x	x
9	(Sub-)Humid plains with very impermeable bedrock	x		
10	(Sub-)Humid mountains with permeable bedrock	x	x	
11	(Sub-)Humid plains/plateaus with impermeable bedrock	x		
12	Very humid plains with impermeable bedrock	x	x	

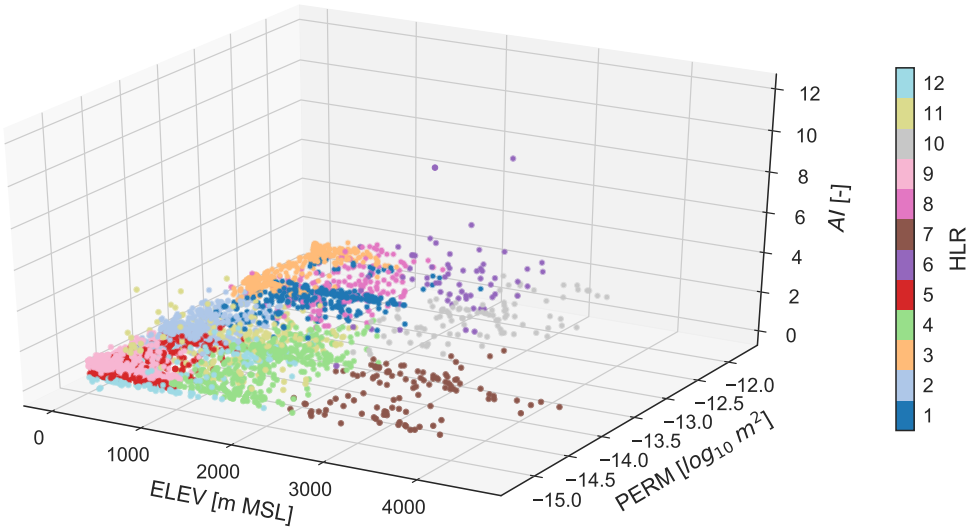


Figure 4.3: 3-D graph of hydrologic landscape regions (HLR). Logarithmic transformation of AI was undone.

Table 4.2: Mean and standard deviation of aridity index (AI), surface elevation ($ELEV$) and permeability of geology ($PERM$) calculated for hydrologic landscape regions (HLR). Logarithmic transformation of AI was undone.

HLR number	AI [-]	$ELEV$ [m MSL]	$PERM$ [log_{10} m^2]
1	0.47 ± 0.06	695 ± 200	-12.7 ± 0.2
2	0.98 ± 0.10	392 ± 103	-13.9 ± 0.2
3	1.01 ± 0.10	214 ± 93	-12.3 ± 0.3
4	0.70 ± 0.08	1431 ± 161	-14.3 ± 0.3
5	0.55 ± 0.05	432 ± 115	-14.7 ± 0.2
6	2.60 ± 0.75	2217 ± 209	-12.7 ± 0.3
7	0.81 ± 0.13	3082 ± 249	-14.4 ± 0.3
8	1.36 ± 0.19	1050 ± 161	-12.7 ± 0.3
9	1.03 ± 0.10	413 ± 112	-14.9 ± 0.1
10	1.00 ± 0.13	2558 ± 216	-12.7 ± 0.3
11	1.59 ± 0.25	1166 ± 224	-14.3 ± 0.3
12	0.27 ± 0.04	787 ± 198	-14.6 ± 0.3

To illustrate how different hydrologic landscape regions may be conceptualized in different hydrologic systems we give several examples in the following:

- (Sub-)Humid plains with permeable bedrock (HLR 2) are characterised by flat topography, a surplus of precipitation over potential evaporation and permeable bedrock. Groundwater recharge is expected to be high. Hence, shallow and deep groundwater are assumed to be important components of the hydro-

logical system.

- Humid low range mountains with impermeable bedrock (HLR 4) have a medium altitude and mountanious landscape morphology, a surplus of precipitation over potential evaporation and impermeable bedrock. Hydrologic flow paths are dominated by overland flow whereas groundwater recharge is less important.
- Humid mountains with impermeable bedrock (HLR 7) are described by high altitude and mountanious landscape morphology, a surplus of precipitation over potential evaporation and impermeable bedrock. Given this geologic setting groundwater recharge is expected to be minimal. Due to the steep terrain the hydrologic flow path is dominated by overland flow whereas groundwater recharge is limited.
- (Sub-)Humid plains with very impermeable bedrock (HLR 9) can be summarised as follows: flat topography, a surplus of precipitation over potential evaporation and very impermeable bedrock. Embedded into this setting groundwater recharge is minimal and the main contributor is overland flow.

4.2 Model Evaluation

Figure 4.4 presents maps of $B_{std-sqrt}$ for all five flow percentiles. For all five flow percentiles the ensemble consistently underestimates the runoff for some parts of southern Siberia and for the central part of the Rocky Mountains. For the latter one, the negative bias is more pronounced for high flows (Fig. 4.4a) compared to the low flows (Fig. 4.4e). Furthermore, the spatial coherent region of underestimation for the high flows reaches up to Alaska. In Japan high flows are slightly overestimated, but for the remaining four flow percentiles the understimation increases towards the low flows. Generally, the negative $B_{std-sqrt}$ appears to be more pronunced for low flows. By contrast, there is no region for which the positive $B_{std-sqrt}$ points for all flow percentiles consistently in one direction. Regarding the moderate low flows and low flows the ensemble shows an overestimation for the East Coast of the USA. Generally, the ensemble performs well for the central USA and the UK. Figure 4.5 exhibits maps of B_{rel} for all five flow percentiles. These maps reveal patterns similar to those in Figure 4.4; the B_{rel} is, however, more accentuated for smaller catchments (Fig. 2.3). This can be observed for the overestimation of moderate high flows and high flows in Western and Southern Russia (Figs. 4.4a, 4.4b, 4.5a, and

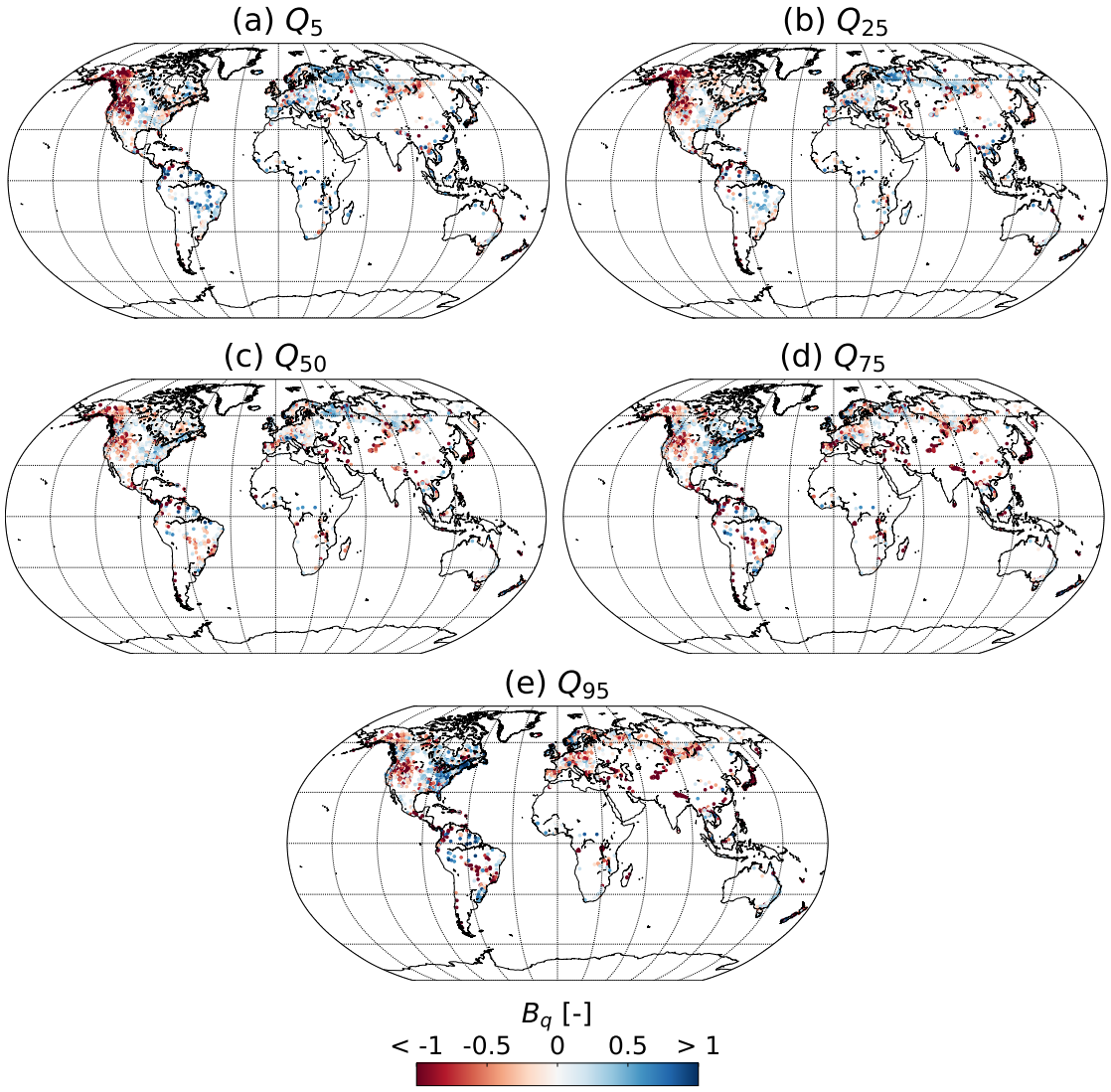


Figure 4.4: Global maps of $B_{std-sqrt}$ for the five flow percentiles (a-e) computed by equation (3.3). Red (blue) indicates a negative (positive) $B_{std-sqrt}$. Hotspots for strong underestimation are found in Alaska (a-e), the Rocky Mountains (a-e), central Asia (d-e), and Japan (c-e), whereas hotspots for strong overestimation are located in the US East Coast (d-e) and western Russia (a-b).

4.5b) and also for low flows of the US East Coast (Figs. 4.4e and 4.5e). Although we computed B_{std}, we found the results not suitable because the data was less normally distributed than $B_{std-sqrt}$. Thus, we decided to exclude them from the statistical analysis and will not show any results. Furthermore, we argue that one standardised bias is sufficient.

Figure 4.6 shows maps of $KGE_{\gamma\beta}$ and its two components γ and β. With respect to γ an overall tendency towards slight underestimation is predominant with only some exceptions (e.g., Central Europe, New Zealand). The picture is more

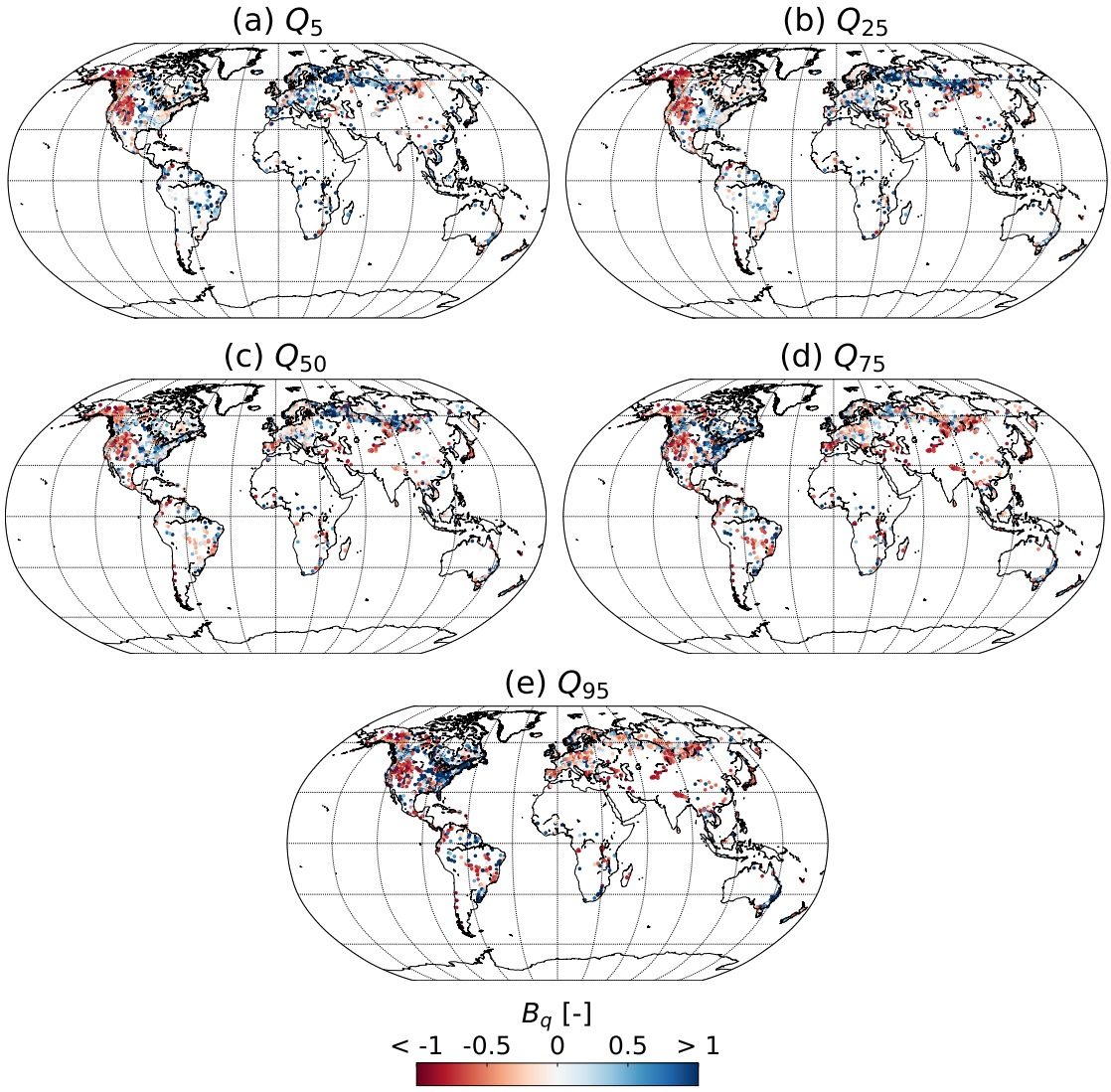


Figure 4.5: Global maps of B_{rel} for the five flow percentiles (a-e) computed by equation (3.1). Maps as Figure 4.4. Hotspots for strong underestimation are found in Alaska (a-e) and the Rocky Mountains (a-e), central Brasil (d-e), central Asia (d-e), and Japan (c-e), whereas hotspots for strong overestimation are located in the US East Coast (d-e), western Russia (a-b), partly in southern Russia (a-b), south-western Australia (e).

complex for β and the pattern resembles those in Figure 4.4 and 4.5, which show an underestimation for Alaska, the Rocky mountains and parts of southern Siberia. Overestimation is mainly found in western and southern Russia. Particularly, the $KGE_{\gamma\beta}$ performs poorly for those regions where β is overestimated and underestimated, respectively. A good performance is found for the US East Coast and the UK.

Figure 4.7 depicts maps of CV for all five flow percentiles which show that the

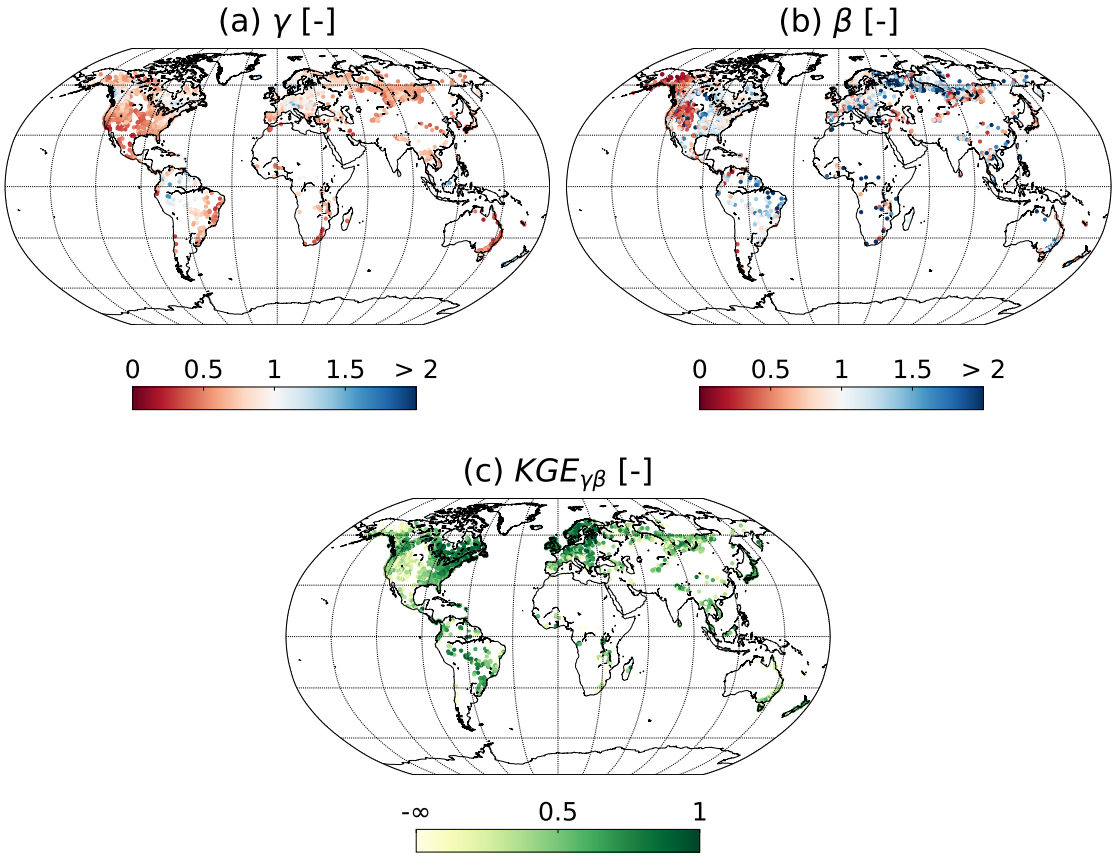


Figure 4.6: Global maps of γ (a), β (b) and $KGE_{\gamma\beta}$ (c) computed by equations (3.4 — 3.6). Blue (red) displays an overestimated (underestimated) γ and β, respectively. Dark green (light green) indicates a good (poor) $KGE_{\gamma\beta}$. A hotspot of poor $KGE_{\gamma\beta}$ is located in the Rocky Mountains.

inter-model disagreement is systematically high for all five flow percentiles in Alaska, the Rocky Mountains and southern Siberia. We found low inter-model disagreement for the US East Coast, the UK, Japan, New Zealand, central and northern Europe, and western Russia; however, for the US East Coast, central and northern Europe the disagreement increases towards the low flows.

After inspecting the distributions of $B_{std-sqrt}$ and B_{rel} for all HLRs, we selected HLR 2, HLR 3, HLR 8, HLR 9, and HLR 11. We call them selected HLRs hereafter. Those HLRs which were not listed are refered to as non-selected HLRs hereafter. We further present the underlying mean, median, standard deviation, skewness, and kurtosis of those distributions in Table 4.3. Figure 4.8 depicts the distributions of $B_{std-sqrt}$, B_{rel}, and CV on all five flow percentiles for the entire dataset and selected HLRs. For the $B_{std-sqrt}$ of the entire dataset, no clear pattern is detectable. The medians are negative, but close to zero. The standard deviation does not vary much

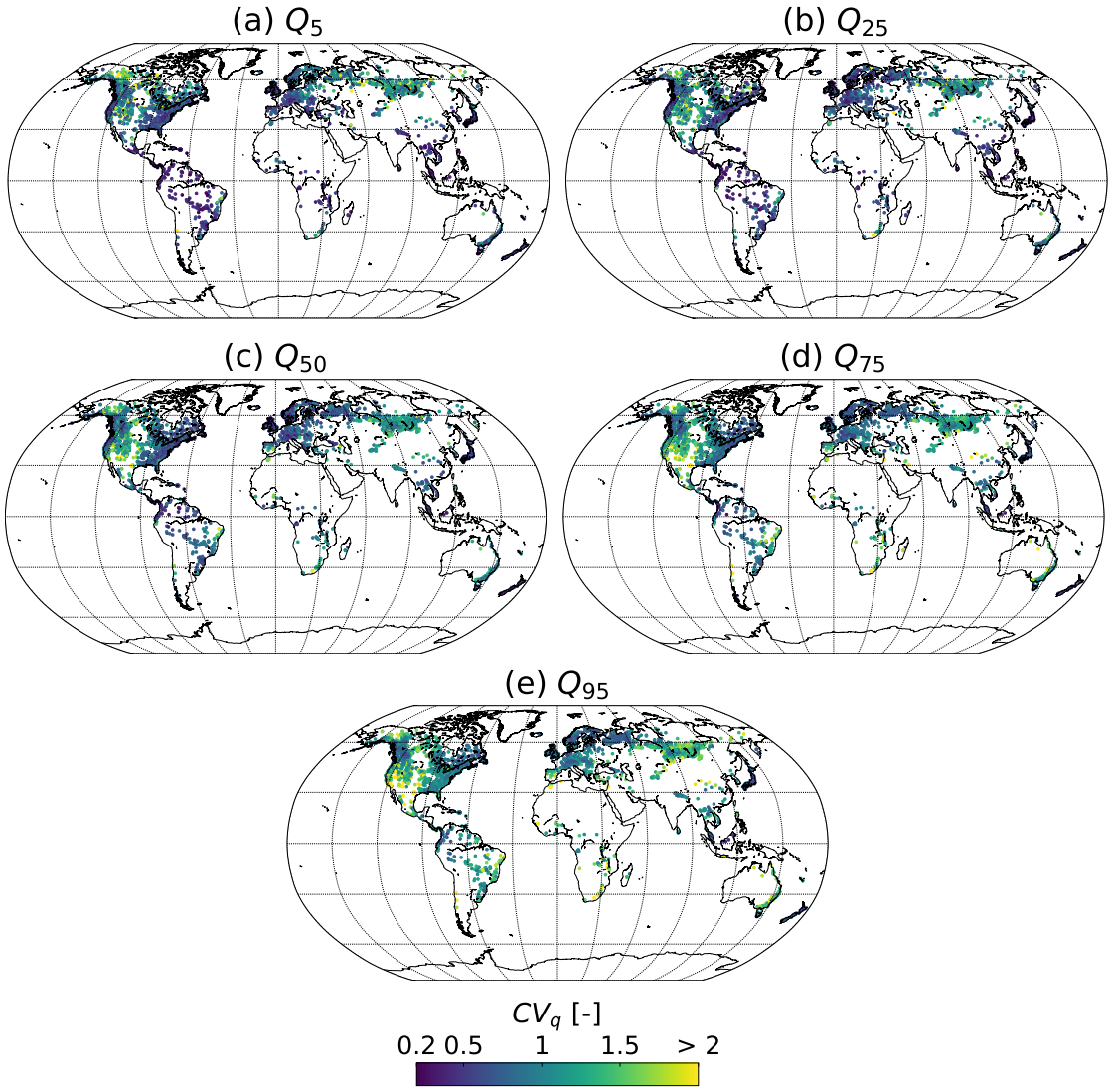


Figure 4.7: Global maps of CV for the five flow percentiles (a-e) computed by equation (3.7). Blue (yellow) depicts low (high) inter-model disagreement. Hotspots of high inter-model disagreement exist for Alaska, the Rocky Mountains and southern Russia.

among the percentiles. This also holds for the slightly negative skewness values (Table 4.3). Similarly, there is also no pattern in B_{rel} of the entire dataset. Likewise medians are negative and close to zero, but in contrast means are positive as well as values for skewness (Table 4.3). When considering the distributions of $B_{std-sqrt}$ and B_{rel} of all selected HLRs two distinct issues can be identified: (i) mean values of $B_{std-sqrt}$ are always less than median values; (ii) mean values of B_{rel} are greater than median values. Common patterns are shared by HLR 2, HLR 3, HLR 8, and HLR 11. In particular, they tend to have positive median values for Q_5 and Q_{25} and negative median values for Q_{75} and Q_{95}. Another common ground for those HLRs

is that $B_{std-sqrt}$ and B_{rel} show a larger spread at Q_{95} than at Q_5. This is reflected by a greater standard deviation (Table 4.3). By contrast, $B_{std-sqrt}$ and B_{rel} in HLR 9 show a tendency towards positive biases for Q_{50}, Q_{75}, and Q_{95}. For Q_5 and Q_{25} this tendency is diminishing.

For the non-selected HLRs we found especially consistent patterns for HLRs with complex topography (HLR 1, HLR 4, HLR 7, HLR 10, and HLR 12). The complex topography manifests, for example, in higher SLO (Fig. A.10k). Their distributions of $B_{std-sqrt}$ and B_{rel} completely point towards a negative bias. For further details we refer to Figure A.11.

The CV rises consistently for the entire dataset and the selected HLRs from high flows to low flows (Figs. 4.8k-4.8o). HLR 8 and HLR 11 appear to have highest CV values (Figs. 4.8m and 4.8o, respectively). For HLR 2, HLR 3, and HLR 9 distributions are similar to those of the entire dataset (Figs. 4.8k, 4.8l, and 4.8n, respectively).

KS statistic and p value derived from the two-sample Kolmogorov-Smirnov-Test between the distribution of the entire dataset and every HLR are given in Table 4.4. At a significance level of 5% (two-tailed p value) most distributions of $B_{std-sqrt}$ and B_{rel} are significantly different from each other. Exceptions exist, for example, in HLR 3 where for B_{rel} Q_{75}, $B_{std-sqrt}$ Q_{95}, and CV Q_{25} distributions are the same.

4.3 Bivariate Regression: Climatic and Physiographic Controls on Model Errors

The R^2 values obtained from the bivariate regression analysis are illustrated in Figure 4.9. Best R^2 values suffice just to describe weak relationships. Moreover, no relationship for the majority of the predictor-predictand pairs is available. For informative predictor-predictand pairs with $R^2 > 0.15$ scatterplots are shown in Figure 4.11. The weak relationships were predominantly nonlinear and/or often characterised by heteroscedasticity (i.e. uneven variance of the data). For the sake of simplicity we placed the predictors into groups of climate, topography (T), land cover (LC), geology, soils (So), and water use (WU). The results unveil that at this level ($R^2 > 0.15$) seven predictors have an influence on the model errors. From those, three relate to climate, one to land cover, one to topography, one to geology, and one to Water use. fS was negatively related to B_{rel} Q_5 in HLR 8 (Fig. 4.11d). Negative relationships were also found for $KGE_{\gamma\beta}$ between P_{si} in HLR 9 and AI in HLR 11,

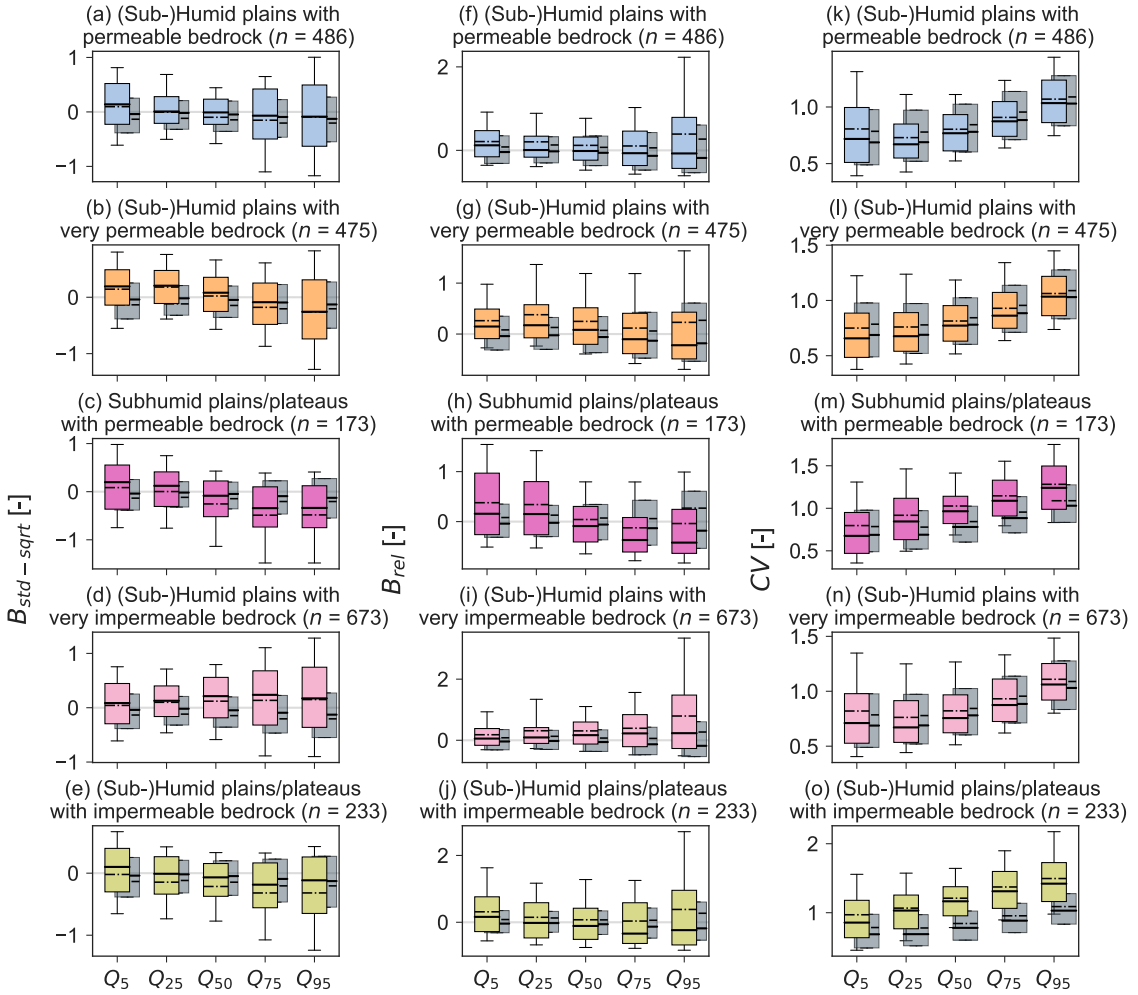


Figure 4.8: Distributions of B_{rel} (a-e), $B_{std-sqrt}$ (f-j), and CV (k-o) on all five flow percentiles for the entire dataset and selected HLRs. n refers to the number of data points. The box plot whiskers range from the 10% to the 90% percentile of the distribution, the box represents the inter-quartile range, solid line depicts the median, and dashed line displays the mean. The dark grey boxes in the background represent the boxes of the entire dataset ($n = 3635$). For catchments locations see Fig. 4.2.

respectively (Figs. 4.11g and 4.11h, respectively). SLO was negatively related to $B_{std-sqrt}$ Q_95 in HLR 8 (Fig. 4.11b). Conversely, the relationship between $CORR$ and $B_{std-sqrt}$ Q_95 in HLR 8 is positive (Fig. 4.11a). Further positive relationships could be shown between AI and B_{rel} Q_5 in HLR 8 (Fig. 4.11c), between fLi_{ss} and B_{rel} Q_{75} in HLR 8 (Fig. 4.11f), and between IRR and B_{rel} Q_5 in HLR 8 (Fig. 4.11e). Besides that, the latter one is characterised by a high heteroscedasticity. The variance is high for low IRR and vice versa.

The results of the bivariate regression describing the predictor-predictand rela-

Table 4.3: Mean, median, standard deviation, skewness, and kurtosis of evaluation metrics for the entire dataset and selected HLRs. n refers to the number of data points. In each subset the mean, median, standard deviation skewness and kurtosis of the distribution of each evaluation metric are shown. Table 4.1 lists the descriptions of the HLRs.

Metric		Mean	Median	Standard deviation	Skewness	Kurtosis		Mean	Median	Standard deviation	Skewness	Kurtosis		Mean	Median	Standard deviation	Skewness	Kurtosis
$B_{std-sqrt}$ Q_5		-0.13	-0.04	0.65	-2.21	15.61		0.1	0.14	0.75	-3.07	26.41		0.14	0.19	0.58	-0.76	2.8
$B_{std-sqrt}$ Q_{25}		-0.12	-0.02	0.6	-3.1	30.39		-0.01	0.01	0.79	-6.82	80.16		0.18	0.2	0.55	-1.12	8.13
$B_{std-sqrt}$ Q_{50}		-0.15	-0.05	0.62	-4.23	50.12		-0.1	-0.01	0.85	-8.56	105.92		0.02	0.08	0.65	-2.74	19.87
$B_{std-sqrt}$ Q_{75}		-0.2	-0.09	0.69	-2.02	12.08		-0.15	-0.07	0.78	-1.62	7.42		-0.18	-0.09	0.79	-2.98	21.63
$B_{std-sqrt}$ Q_{95}		-0.2	-0.13	0.76	-1.13	4.03		-0.09	-0.09	0.87	-0.4	0.64		-0.26	-0.26	0.86	-0.67	1.67
B_{rel} Q_5	**All (n = 3635)**	0.08	-0.04	0.59	1.15	1.54	**HLR 2 (n = 486)**	0.21	0.12	0.53	1.08	1.54	**HLR 3 (n =475)**	0.26	0.15	0.52	1.14	1.5
B_{rel} Q_{25}		0.13	-0.02	0.71	1.78	4.07		0.2	0.01	0.7	2.16	5.66		0.38	0.17	0.71	1.75	3.47
B_{rel} Q_{50}		0.07	-0.06	0.65	1.53	3.5		0.12	-0.02	0.6	1.62	3.84		0.25	0.08	0.66	1.39	2.3
B_{rel} Q_{75}		0.06	-0.13	0.75	1.49	2.97		0.1	-0.07	0.67	1.35	2.55		0.12	-0.1	0.76	1.74	3.63
B_{rel} Q_{95}		0.27	-0.18	1.22	1.8	3.03		0.39	-0.07	1.22	1.67	2.33		0.23	-0.22	1.19	2.07	4.14
$KGE_{\gamma\beta}$		0.51	0.57	0.32	-1.79	5.81		0.57	0.63	0.33	-1.97	5.48		0.52	0.57	0.28	-1.97	5.8
CV Q_5		0.79	0.69	0.42	1.53	3.33		0.81	0.72	0.39	1.19	1.56		0.75	0.66	0.4	1.88	5.05
CV Q_{25}		0.78	0.69	0.38	1.23	2.02		0.73	0.67	0.3	1.4	2.87		0.76	0.68	0.34	1.47	2.96
CV Q_{50}		0.84	0.78	0.34	1.14	2.05		0.8	0.77	0.26	1.28	2.8		0.81	0.77	0.27	0.85	1.19
CV Q_{75}		0.96	0.88	0.35	1.17	2.32		0.91	0.87	0.24	0.84	0.98		0.93	0.86	0.26	0.97	1.05
CV Q_{95}		1.09	1.03	0.36	1.22	2.42		1.07	1.03	0.27	0.65	0.31		1.06	1.03	0.27	0.79	0.9
$B_{std-sqrt}$ Q_5		0.08	0.2	0.78	-1.35	4.47		0.04	0.09	0.73	-6.15	94.18		-0.02	0.1	0.94	-7.77	89.91
$B_{std-sqrt}$ Q_{25}		0	0.12	0.75	-1.93	7.75		0.1	0.13	0.62	-2.51	21.15		-0.15	-0.01	0.95	-8.79	103.86
$B_{std-sqrt}$ Q_{50}		-0.25	-0.08	0.78	-2.18	7.5		0.12	0.21	0.69	-2.24	11.9		-0.21	-0.07	0.95	-9.47	118.2
$B_{std-sqrt}$ Q_{75}		-0.49	-0.34	0.86	-1.4	2.54		0.14	0.24	0.86	-1.47	6.41		-0.32	-0.19	0.92	-6.53	67.89
$B_{std-sqrt}$ Q_{95}		-0.48	-0.34	0.89	-1.75	4.26		0.15	0.17	0.93	-0.92	2.45		-0.32	-0.12	0.88	-3.75	27.41
B_{rel} Q_5	**HLR 8 (n =173)**	0.38	0.16	0.8	0.65	-0.54	**HLR 9 (n = 673)**	0.18	0.05	0.54	1.26	2.1	**HLR 11 (n = 233)**	0.31	0.16	0.79	0.75	-0.26
B_{rel} Q_{25}		0.34	0.15	0.85	1.24	1.84		0.31	0.09	0.76	1.97	4.04		0.15	-0.02	0.75	0.83	0.15
B_{rel} Q_{50}		0.04	-0.09	0.7	1.68	3.84		0.31	0.17	0.67	1.42	3.15		0.08	-0.11	0.8	1.22	1.23
B_{rel} Q_{75}		-0.12	-0.37	0.76	2.27	7.34		0.39	0.22	0.83	1.15	1.58		0.04	-0.34	0.91	1.51	2.42
B_{rel} Q_{95}		-0.04	-0.42	1.01	2.28	5.83		0.79	0.23	1.47	1.26	0.7		0.38	-0.24	1.49	1.6	1.93
$KGE_{\gamma\beta}$		0.38	0.47	0.36	-1.54	2.79		0.53	0.62	0.37	-2.43	8.48		0.29	0.33	0.35	-1.31	2.99
CV Q_5		0.8	0.68	0.47	1.87	4.36		0.82	0.71	0.44	1.8	4.18		0.97	0.86	0.5	1.37	2.21
CV Q_{25}		0.92	0.85	0.38	0.75	0.32		0.76	0.67	0.34	1.51	3.74		1.07	1.03	0.4	0.82	0.92
CV Q_{50}		1.03	0.97	0.33	1.24	2.13		0.82	0.76	0.3	1.06	1.14		1.21	1.17	0.38	1.39	3.34
CV Q_{75}		1.15	1.09	0.33	1.12	2.11		0.93	0.87	0.28	0.89	0.93		1.37	1.31	0.45	1.11	1.44
CV Q_{95}		1.28	1.24	0.38	0.77	0.96		1.11	1.06	0.29	1.2	2.68		1.5	1.42	0.47	0.93	0.66

Table 4.4: Kolmogorov-Smirnov statistic for distributions of the evaluation metrics between the entire dataset and HLRs

	HLR 1		**HLR 2**		**HLR 3**		**HLR 4**		**HLR 5**		**HLR 6**	
	KS statistic	p value	KS statistic	p value	KS statistic	p value	KS statistic	p value	KS statistic	p value	KS statistic	p value
$B_{std-sqrt}$ Q_5	0.191	1.77e-09	0.174	8.02e-12	0.21	1.21e-16	0.32	1.95e-35	0.132	1.51e-06	0.228	1.75e-03
$B_{std-sqrt}$ Q_{25}	0.295	4.48e-22	0.091	1.50e-03	0.257	8.67e-25	0.289	5.33e-29	0.12	1.93e-05	0.275	6.93e-05
$B_{std-sqrt}$ Q_{50}	0.27	1.62e-18	0.088	2.26e-03	0.16	6.78e-10	0.24	4.78e-20	0.161	1.69e-09	0.278	5.49e-05
$B_{std-sqrt}$ Q_{75}	0.192	1.48e-09	0.128	1.35e-06	0.049	0.259	0.195	2.30e-13	0.184	2.88e-12	0.206	6.39e-03
$B_{std-sqrt}$ Q_{95}	0.169	1.69e-07	0.115	2.12e-05	0.095	1.00e-03	0.214	5.08e-16	0.176	2.64e-11	0.166	4.80e-02
B_{rel} Q_5	0.137	4.33e-05	0.153	2.68e-09	0.206	3.71e-16	0.265	2.05e-24	0.124	7.99e-06	0.161	0.06
B_{rel} Q_{25}	0.229	1.71e-13	0.116	1.73e-05	0.219	4.60e-18	0.223	2.13e-17	0.11	1.11e-04	0.228	1.73e-03
B_{rel} Q_{50}	0.221	1.36e-12	0.101	2.81e-04	0.151	6.80e-09	0.19	8.61e-13	0.086	4.94e-03	0.281	4.48e-05
B_{rel} Q_{75}	0.171	1.11e-07	0.09	1.65e-03	0.089	2.29e-03	0.152	2.55e-08	0.125	6.97e-06	0.259	2.22e-04
B_{rel} Q_{95}	0.137	4.41e-05	0.108	8.42e-05	0.055	0.148	0.163	1.53e-09	0.13	2.53e-06	0.276	6.42e-05
$KGe_{\gamma\beta}$	0.147	7.84e-06	0.116	1.80e-05	0.092	1.46e-03	0.101	7.16e-04	0.272	2.02e-26	0.585	1.18e-20
CV Q_5	0.308	4.72e-24	0.047	0.295	0.071	2.85e-02	0.131	2.91e-06	0.102	4.87e-04	0.432	1.83e-11
CV Q_{25}	0.351	3.73e-31	0.099	4.15e-04	0.055	0.151	0.138	5.87e-07	0.214	1.69e-16	0.653	1.24e-25
CV Q_{50}	0.361	7.19e-33	0.095	7.89e-04	0.069	3.43e-02	0.1	8.01e-04	0.265	4.64e-25	0.722	3.21e-31
CV Q_{75}	0.244	3.41e-15	0.099	4.37e-04	0.079	1.02e-02	0.084	7.84e-03	0.327	8.45e-38	0.723	2.66e-31
CV Q_{95}	0.276	2.39e-19	0.065	4.85e-02	0.076	1.50e-02	0.061	0.111	0.376	7.70e-50	0.675	2.70e-27
	HLR 7		**HLR 8**		**HLR 9**		**HLR 10**		**HLR 11**		**HLR 12**	
$B_{std-sqrt}$ Q_5	0.404	1.16e-15	0.247	2.25e-09	0.147	2.90e-11	0.399	1.19e-14	0.152	7.04e-05	0.517	2.81e-30
$B_{std-sqrt}$ Q_{25}	0.234	1.57e-05	0.207	1.15e-06	0.184	2.71e-17	0.459	2.81e-19	0.062	0.359	0.495	9.12e-28
$B_{std-sqrt}$ Q_{50}	0.34	3.14e-11	0.084	0.182	0.285	6.38e-41	0.424	1.90e-16	0.07	0.230	0.372	7.28e-16
$B_{std-sqrt}$ Q_{75}	0.349	8.32e-12	0.185	2.07e-05	0.28	2.17e-39	0.374	6.34e-13	0.079	0.128	0.217	1.12e-05
$B_{std-sqrt}$ Q_{95}	0.344	1.65e-11	0.178	4.91e-05	0.237	2.01e-28	0.359	6.29e-12	0.081	0.11	0.196	1.01e-04
B_{rel} Q_5	0.324	3.18e-10	0.212	5.03e-07	0.155	1.91e-12	0.377	4.29e-13	0.169	5.82e-06	0.408	5.29e-19
B_{rel} Q_{25}	0.201	3.25e-04	0.182	2.81e-05	0.186	1.38e-17	0.443	6.41e-18	0.109	9.82e-03	0.369	1.36e-15
B_{rel} Q_{50}	0.31	1.96e-09	0.057	0.653	0.194	3.55e-19	0.386	9.81e-14	0.113	6.50e-03	0.218	1.02e-05
B_{rel} Q_{75}	0.348	9.79e-12	0.184	2.13e-05	0.207	1.06e-21	0.356	9.91e-12	0.146	1.43e-04	0.091	0.233
B_{rel} Q_{95}	0.402	1.56e-15	0.172	9.42e-05	0.199	3.29e-20	0.387	8.73e-14	0.125	1.99e-03	0.107	0.104
$KGe_{\gamma\beta}$	0.287	3.85e-08	0.181	3.07e-05	0.107	4.22e-06	0.398	1.44e-14	0.38	2.17e-28	0.113	0.075
CV Q_5	0.2	3.80e-04	0.046	0.875	0.055	0.063	0.227	4.90e-05	0.205	1.47e-08	0.272	1.09e-08
CV Q_{25}	0.412	2.91e-16	0.198	3.69e-06	0.057	0.051	0.398	1.37e-14	0.364	4.97e-26	0.307	5.87e-11
CV Q_{50}	0.462	2.06e-20	0.323	1.05e-15	0.055	0.062	0.433	3.69e-17	0.468	8.81e-43	0.419	5.74e-20
CV Q_{75}	0.413	2.23e-16	0.299	1.58e-13	0.042	0.26	0.438	1.49e-17	0.456	1.28e-40	0.379	2.01e-16
CV Q_{95}	0.234	1.55e-05	0.259	2.96e-10	0.129	9.26e-09	0.326	6.45e-10	0.436	2.83e-37	0.29	8.19e-10

tionship for the non-selected HLRs will not be shown (see Sect. 5.1).

In contrast to the model errors, informative relationships found in the inter-model disagreement are characterised by greater strength (Fig. A.4). Important predictors are predominantly related to climate and land cover. The latter one restricts to Q_5. In the appendix, we present informative relationships ($R^2 > 0.3$) between climatic and physiographic characteristics and inter-model disagreement (see Fig. A.6 for further details).

Concerning the scaling effect low R^2 values derived from the regression analysis between the evaluation metrics and the catchment size for the selected HLRs suggesting that there is no systematic error (Table A.1).

4.4 Random Forest: Climatic and Physiographic Controls on Model Errors

Figure 4.10 shows ranks of importance of climatic and physiographic characteristics found by using RF to predict the model errors. Those results for the selected HLRs are mainly in agreement with those found by the bivariate regression. For the sake of simplicity we placed the predictors into groups of climate, topography (T), land cover (LC), geology, soils (So), and water use (WU). From the entire dataset (Fig. 4.10a) we obtained that especially SLO and fS have high relevance for biases of high flows. Topography-related predcitors exhibited a high importance over all flow percentiles in general. Moreover, it is clearly discernible that P_{si} and $PERM$ gain importance for biases of Q_{75} and Q_{95} while the importance of AI is restricted to Q_{25} and Q_5. However, high ranks for the selected HLRs (Figs. 4.10b-4.10f) are hetergeneously distributed among the predictor groups. Hence, the general picture drawn by those findings is more complex than that resulting from the bivariate regression. The results of the RF approach are only described for HLR 3, HLR 8, and HLR 9 here. The reason for that is the shift from positive biases of Q_5 to negative biases of Q_{95} is most distinct in HLR 3 and HLR 8 (Figs. 4.8c, 4.8d, 4.8g, and 4.8h, respectively). By contrast, biases of Q_{75} and Q_{95} in HLR 9 were predominantely positive. We found distinct different patterns between HLR 3, HLR 8, and HLR 9:

- (Sub-)Humid plains with very permeable bedrock (HLR 3): Regarding the predictor importances (Fig. 4.10c) there is clearly a difference between the three metrics describing the model error. Topography-related and water use-related

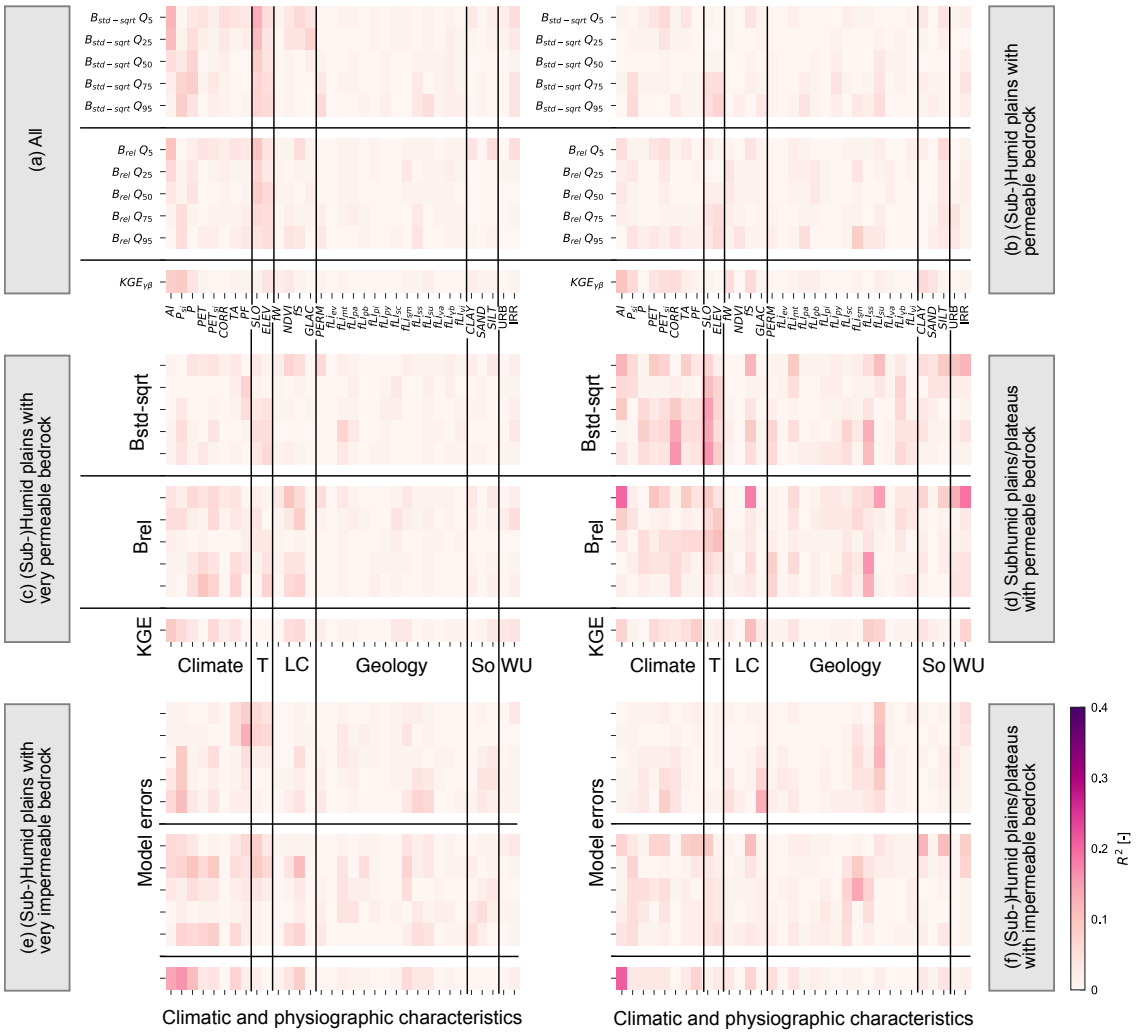


Figure 4.9: Coefficients of determination (R^2) of bivariate (non-)linear regression for $B_{std-sqrt}$, B_{rel} and $KGE_{\gamma\beta}$. Heatmaps of the R^2 are depicted for entire data (a) and for selected HLRs (b-f). Purple (white) indicates moderate (no) relationship. Abbreviations referring to: T, Topography; So, Soils; WU, Water use.

predictors are relatively unimportant for $B_{std-sqrt}$ Q_5. Important predictors for $B_{std-sqrt}$ Q_{25}, $B_{std-sqrt}$ Q_{50}, and $B_{std-sqrt}$ Q_{95} are related to climate, topography, land cover, and soils. For $B_{std-sqrt}$ Q_{75} predictors related to soils are less important. Instead, water use and geology provide important predictors. For B_{rel} important predictors mainly concentrate on climate and land cover except for Q_{50} and Q_{75} where geology comes in addition. Furthermore, relevant predictors for Q_{75} are complemented by topography. For $KGE_{\gamma\beta}$ important predictors are limited to climate and land cover.

- Subhumid plains/plateaus with permeable bedrock (HLR 8): Predictor importances found for $B_{std-sqrt}$, B_{rel}, and $KGE_{\gamma\beta}$ are mainly in agreement (Fig.

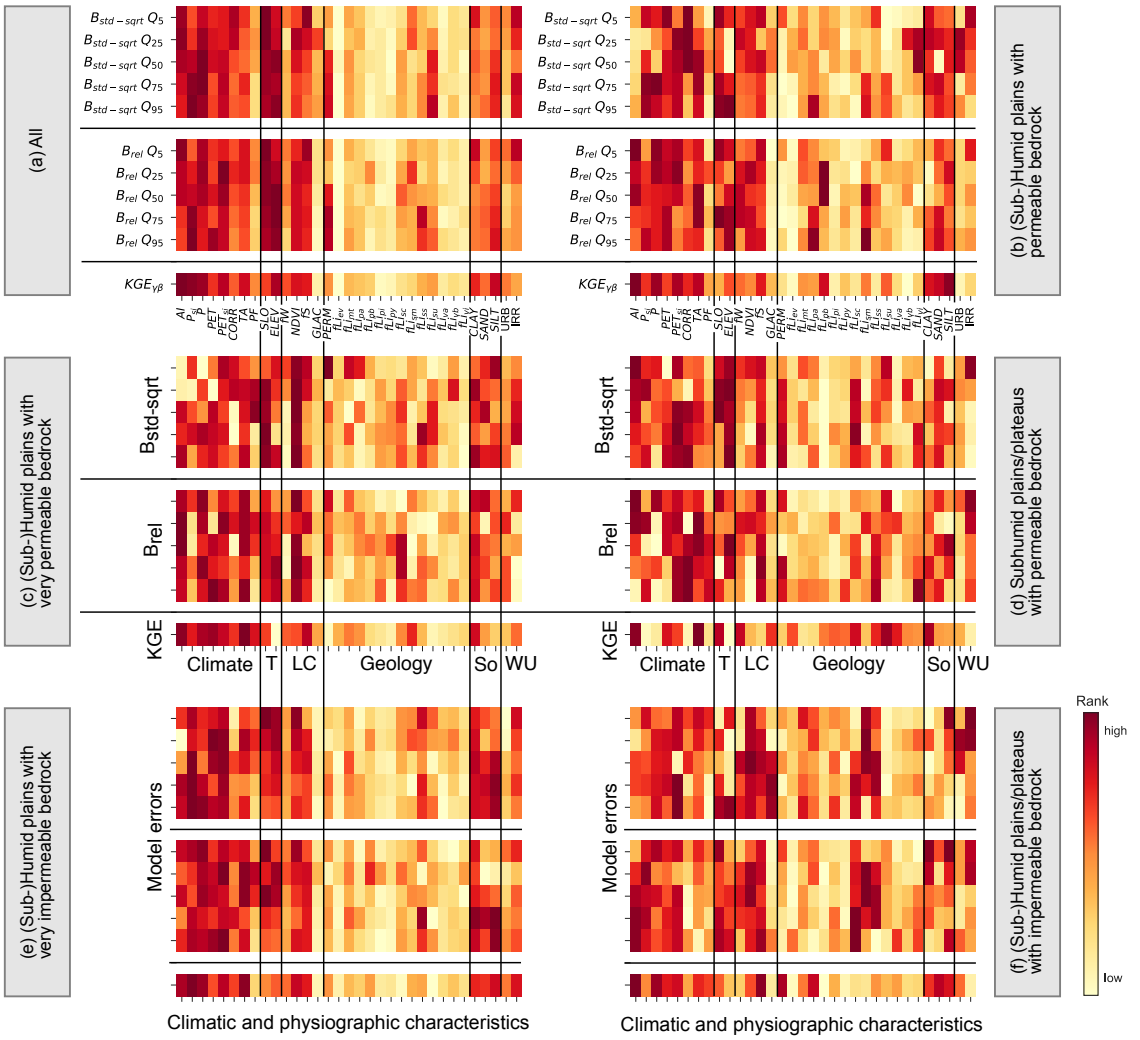


Figure 4.10: Ranks of permutation importance of random forest for $B_{std-sqrt}$, B_{rel} and $KGE_{\gamma\beta}$. Heatmaps of the ranks are depicted for entire data (a) and for selected HLRs (b-f). Red (yellow) displays high (low) ranks. Abbreviations referring to: T, Topography; So, Soils; WU, Water use.

4.10d). This agreement between the two biases is present for Q_5, Q_{75}, and Q_{95}, but not for Q_{25} and Q_{50}. For $B_{std-sqrt}$ Q_{25} climate-related, topography-related, geology-related, and soil-related predictors are relevant while for B_{rel} Q_{25} climate-related, land cover-related, and water use-related predictors are important. Climate-related and topography-related predictors are relevant for both $B_{std-sqrt}$ Q_{50} and B_{rel} Q_{50}. Predictors related to land cover and soils come in addition for B_{rel} Q_{50}, and for $B_{std-sqrt}$ Q_{50} geology provides further important predictors. Regarding $KGE_{\gamma\beta}$ every predictor group provides somewhat relevant predictors except water use.

- (Sub-)Humid plains with very impermeable bedrock (HLR 9): Important pre-

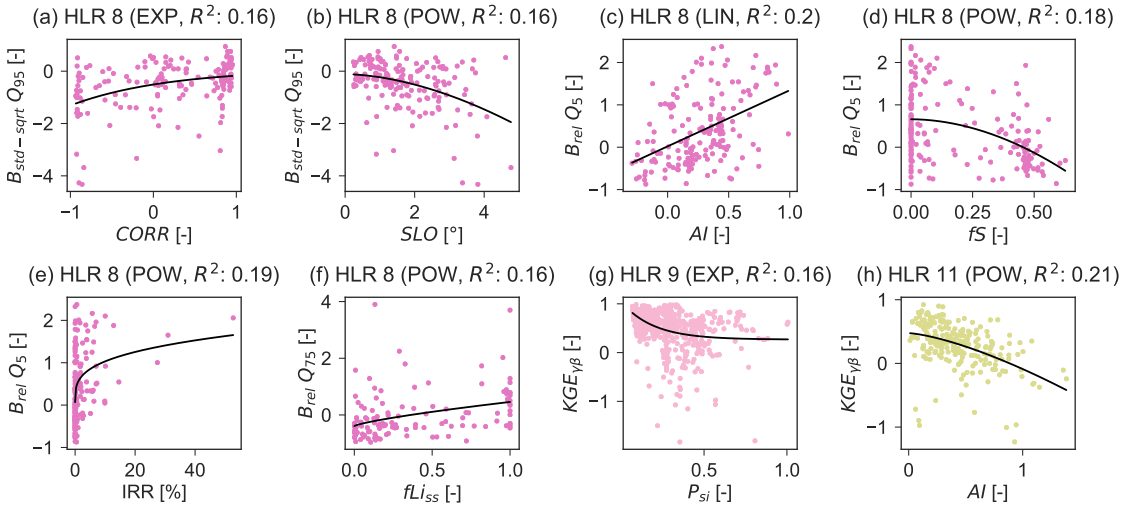


Figure 4.11: Scatterplots of climatic and physiographic characteristics (along the x-axis) versus $B_{std-sqrt}$ and B_{rel} (along the y-axis), including the best-fit regression. Scatterplots are shown for informative predictor-predictand pairs with $R^2 > 0.15$. Each data point represents a catchment. Abbreviations referring to the HLR (for description see Table 4.1) and to the type of the best-fit regression function: EXP, exponential; LIN, linear; LOG, logarithmic; and POW, power.

dictors of $B_{std-sqrt}$ and B_{rel} were found to be in agreement for Q_5 and Q_{95}. For Q_{95} these were associated with climate and soils, whereas for Q_5 these predictors concern climate, topography, and land cover. These three relevant predictor groups are in agreement for $B_{std-sqrt}$ and B_{rel} at Q_{25}. Similar pattern exist for $B_{std-sqrt}$ and B_{rel} at Q_{50} except that for $B_{std-sqrt}$ Q_{50} soils gained further relevance. For biases of Q_{75} and Q_{95} climate and soils reached high importances except for B_{rel} Q_{75} where geology added. For $KGE_{\gamma\beta}$ important predictors are found within climate and land cover.

Predictor importances of HLR 2 and HLR 11 are not described here. Instead, we refer to Figures 4.10b and 4.10f. Heatmaps with predictor importances of non-selected HLRs will not be shown (see Sect. 5.1).

Patterns on the importance of climatic and physiographic characteristics found for the inter-model disagreement (Fig. A.5) emphasize high importance of climate-related predictors over all five flow percentiles. Among those especially AI, P, and PET are associated with higher ranks. At Q_5 further important predictors completing climate-related predictors belong to land cover where fS is highly relevant. The patterns at Q_{25}, Q_{50}, and Q_{75} are characterised by a high spottiness. For example, the at Q_{50} highly ranked predictors are not only related to climate, but also to land cover and geology (Figs. A.5c, A.5d, and A.5f, respectively). At Q_{25} and Q_{75}

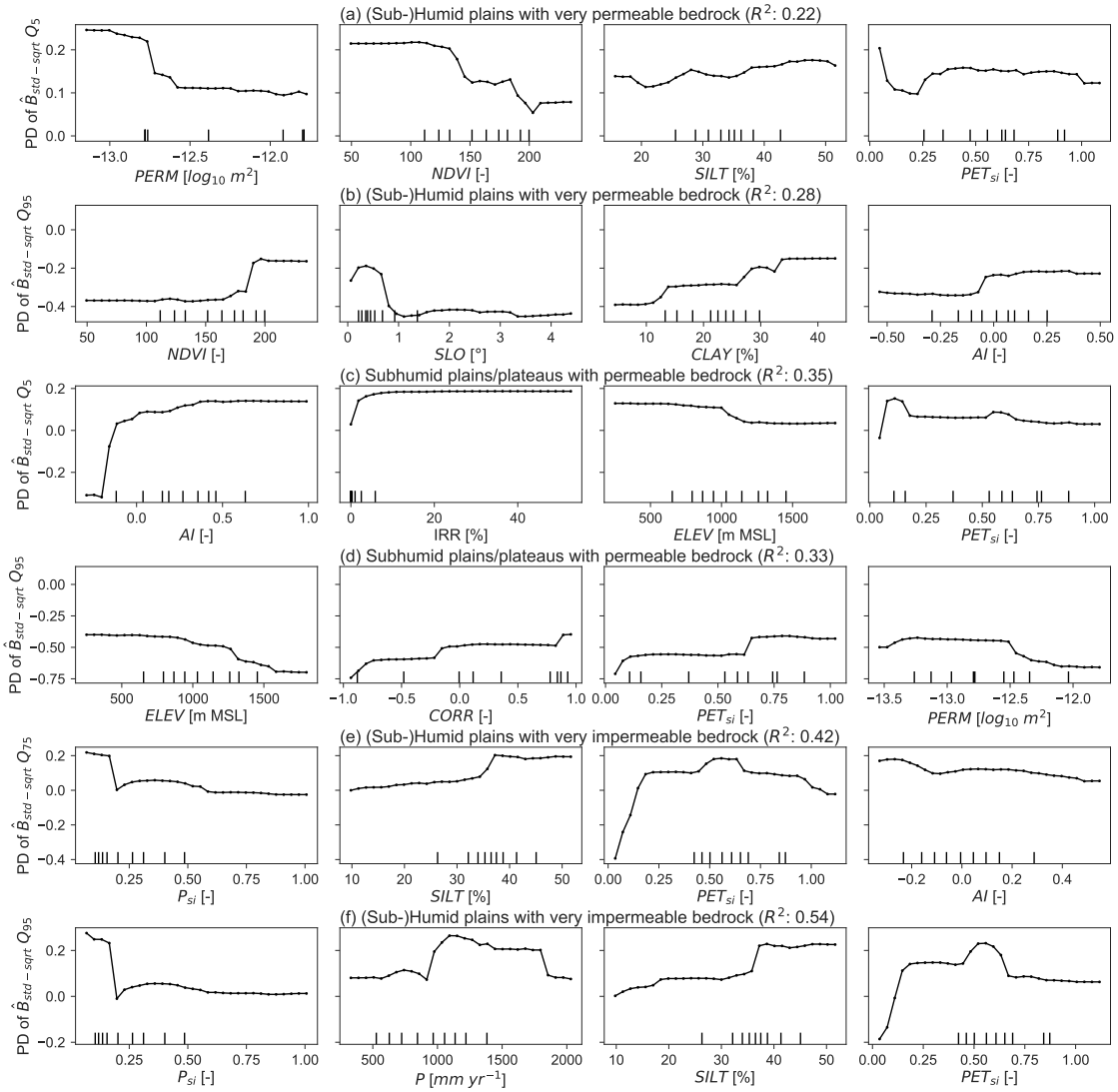


Figure 4.12: Single-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis) versus partial dependence of $\hat{B}_{std-sqrt}$ Q_5, $\hat{B}_{std-sqrt}$ Q_{75}, and $\hat{B}_{std-sqrt}$ Q_{95} (along the z-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). The hash marks at the base of the plots delineate deciles of the corresponding predictor variable. R^2 exhibits the out-of-bag accuracy. The hat ($\hat{}$) denotes the predicted metric by the RF. For abbreviations on x-axis and y-axis see Table 2.2.

important predictors are mainly related to climate, but predictors related to land cover partly appear to be relevant as well (Figs. A.5b and A.5c, respectively).

In Figure 4.12, we present the single-variable partial dependence on $\hat{B}_{std-sqrt}$ Q_5 and $\hat{B}_{std-sqrt}$ Q_{95} in HLR 3 and HLR 8, and single-variable partial dependence on $\hat{B}_{std-sqrt}$ Q_{75} and $\hat{B}_{std-sqrt}$ Q_{95} in HLR 9. These are shown for the four highest ranked predictors. Notice that results of partial dependence will only be described

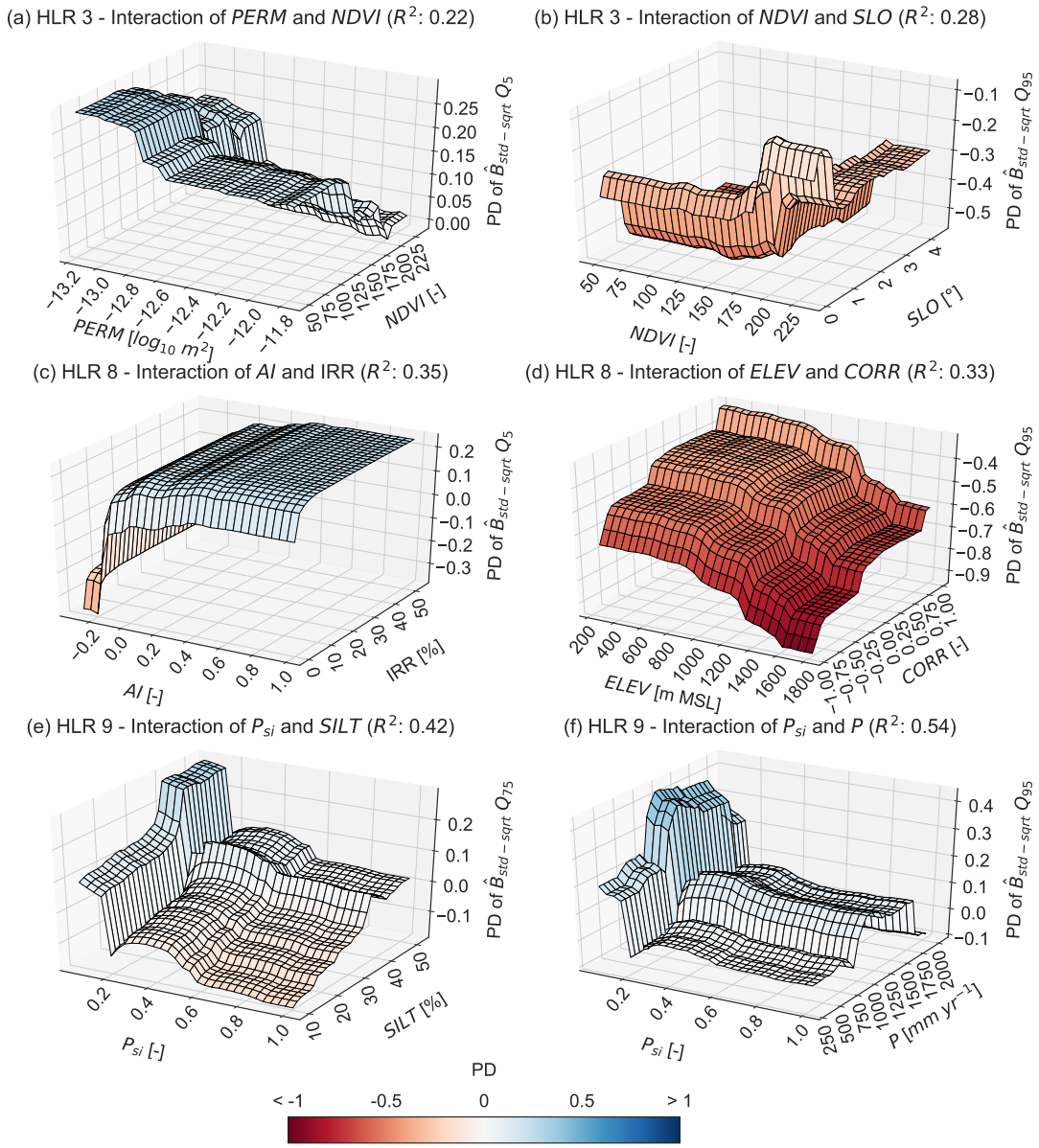


Figure 4.13: Two-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis and y-axis) versus partial dependence of $\hat{B}_{std-sqrt}$ Q_5, $\hat{B}_{std-sqrt}$ Q_{75}, and $\hat{B}_{std-sqrt}$ Q_{95} (along the z-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). R^2 exhibits the out-of-bag accuracy. The hat (ˆ) denotes the predicted metric by the RF. For abbreviations on x-axis see Table 2.2 and for descriptions of HLRs see Table 4.1.

for those cases in which data density is sufficient. Data density is indicated by deciles (see Section 3.5). Since different important predictors are identified for $B_{std-sqrt}$ and B_{rel} we show only the results of $B_{std-sqrt}$ here. For the same partial dependencies as in Figure 4.12 but for B_{rel} we refer to Figures A.2 and A.3. The relationships were predominantly nonlinear and/or often characterised by non-monotony.

In HLR 3 the partial dependence of $\hat{B}_{sqrt}\ Q_5$ (Fig. 4.12a) on $PERM$ is decreasing monotonically. The partial dependence on $NDVI$ also reveals a monotonic decrease. We recall that $NDVI$ ranges from 0 to 255 (see Section 2.4). This is opposed by a monotonic increase existing for partial dependence on $SILT$ and PET_{si}. However, relationships derived from the partial dependence on Q_{95} (Fig. 4.12b) show positive relationship for $CLAY$, $NDVI$, and AI. $NDVI$ exhibits a sudden increase at 180. A sudden increase was also for AI at -0.1. The relationship of SLO is characterised by negative monotonic shape partial dependence decreasing between 0.5 and 1.5.

The partial dependence of $\hat{B}_{sqrt}\ Q_5$ in HLR 8 reveals a slight positive relationship for AI and a strong positive relationship for IRR (Fig. 4.12c). Slight negative relationships are present for $ELEV$ and PET_{si} (Fig. 4.12c). Conversely, PET_{si} behaves in the opposite way for the partial dependence on Q_{95} and increases monotonically (Fig. 4.12d). This is also observed on $CORR$ (Fig. 4.12d). Monotonic decrease exists for partial dependence both on $ELEV$ and $PERM$ (Fig. 4.12d).

In HLR 9 there is a strong positive monotonic partial dependence apparent for P_{si} for $\hat{B}_{std-sqrt}\ Q_{75}$ and $\hat{B}_{std-sqrt}\ Q_{95}$ (Figs. 4.12e and 4.12f). For $SILT$ a monotonic increase was found (Figs. 4.12e and 4.12f). PET_{si} have a non-monotonic partial dependence. Slight negative partial dependence is exhibited on AI for $\hat{B}_{std-sqrt}\ Q_{95}$ (Fig. 4.12e). Partial dependence on P for $\hat{B}_{std-sqrt}\ Q_{95}$ changes its slope several times, but generally displays a positive relationship (Fig. 4.12f).

Corresponding to the single-variable partial dependencies the two-variable partial dependencies of the two highest ranked predictors are illustrated by Figure 4.13. The interaction surfaces are often characterised by high complexity. A flat shape implies that the predicted metric is independent of the two variables considered. In case the surface is inclined towards one variable, then the partial dependence relies just on this variable. If the inclination is directed towards both variables an interaction between the two variables exists. We found strong two-sided interaction just for $\hat{B}_{std-sqrt}\ Q_{95}$ in HLR 8 between $ELEV$ and $CORR$ (Fig. 4.13d). Concerning the remaining two-variable partial dependencies interaction restrain to those areas where non-constant partial dependencies intersect (Figs. 4.13a, 4.13b, 4.13c, 4.13e, and 4.13f, respectively).

The R^2 values of RF which give evidence about the model accuracy range from poor to moderate (Table A.2).

5 Discussion

5.1 Model Evaluation

The model evaluation based on the entire dataset did not provide any interesting insights. On the contrary, assessing the ensemble performance from the perspective of mean and median of $B_{std-sqrt}$ and B_{rel} would suggest that the models have somehow slight biases (Fig. 4.8, Table 4.3). Thereby, it remains invisible where exactly the ensemble performs poorly. The partitioning of the catchments into subsets highlighted those catchments for which the ensemble exhibited significant deficiencies in capturing long-term runoff trends at five different flow percentiles. Results of the two-sample Kolmogorov-Smirnov-Test demonstrated that distributions of biases of selected HLRs were mostly significantly different from distributions of the entire dataset (Table 4.4). This strongly underpinned our strategy to divide data into subsets. By comparing $B_{std-sqrt}$ and B_{rel} (Figs. 4.8 and A.11, Table 4.3) the manifested deficiencies are threefold: (i) consistent tendency of runoff underestimation for catchments with complex topography (e.g., HLRs with higher SLO; Fig. A.10k); (ii) consistent tendency to overestimate runoff for Q_5 and Q_{25} as well as to underestimate runoff for Q_{75} and Q_{95} for (sub-)humid catchments with flat topography (e.g., HLRs with lower SLO; Fig. A.9k). It is conceivable that models release too much of the precipitation too quickly, which would explain the overestimation of Q_5. As a consequence, less water is stored in soils and aquifers which lead to an underestimation of Q_{95}. Since models included in the ensemble account for a closed water balance (Schellekens et al. 2017), one can deduce that the error is attributed to the model structure (e.g., storage routine). In this respect, it would be interesting to see how the distributions for the total biases might look like (i.e. bias is calculated between entire simulated and observed time series); (iii) tendency to overestimate runoff for Q_{75} and Q_{95} for (sub-)humid catchments with flat topography and very impermeable bedrock. The overestimation might be due to wrong parameterisation of the storage (e.g., soil storage and/or groundwater storage). An oversizing of the storage volume might lead to ongoing release of water although the storage already has to be much more depleted.

Due to the presence of a high spatial correlation in the patterns of $B_{std-sqrt}$ and B_{rel} (cf. Figs. 4.4, 4.5, and 4.2) errors may originate from the WFDEI P data. Beck et al. (2017a), for example, have demonstrated for the conterminous USA that model errors and P bias are correlated moderately strong and, thus, propagate into the models. They argue that the biases in the WFDEI P data are present because the forcing does not properly account for orographic effects. Similarly, Gudmundsson et al. (2012b) pointed towards biased forcing data that is possibly attributed to the model error. This suggests that simulated runoff is highly sensitive to the forcing (e.g., Sperna Weiland et al. 2015). Moreover, the P bias may differ substantially for different forcings (e.g., Materia et al. 2010). This makes it complicated to answer the question of whether the error is attributed to model parameterisation or structural errors (e.g., storage routine), or whether the spurious runoff estimates are solely caused by the forcing. To overcome this ambiguity and enhance a process-based evaluation two issues have to be clarified: (i) More effort should be devoted to the global evaluation of P biases in the forcing (e.g., Beck et al. (2017a)); (ii) Catchments which are affected by forcing errors might be excluded from the model evaluation. The evaluation of P errors was beyond the scope of this study. However, we propose using the independent MSWEP P data (see Table 2.2) for profound assessment. Currently, this is the most accurate global-scale P dataset (Beck et al. 2017a). In contrast to the WFDEI dataset they corrected P for gauge under-catch and orographic effects.

One remarkable result of the model evaluation is that, independently from dataset being considered, the inter-model disagreement increases from Q_5 to Q_{95} (Fig. 4.8, Table 4.3). The larger spread for low flows is in line with the findings of Gudmundsson et al. (2012b). A reason for the higher inter-model disagreement is the associated uncertainty in mathematical representations for low flows (Gudmundsson et al. 2012b). Although the models in the ensemble incorporate smiliar processes (see Table 2.1) the general inter-model disagreement for the five flow percentiles might also be explained by different parameterisation. The models use a wide range of data products for setup. Even though models may have the same data sources for parameterisation different processing and interpretation of the mapped values may result in different model parameters (Gudmundsson et al. 2012b).

Several studies have shown that multi-model ensembles lead to an improved performance (e.g., Beck et al. 2017a; Gudmundsson et al. 2012b; Materia et al. 2010); we proved its suitabilty for multi-model evaluation. Using a multi-model ensemble is less cumbersome since the evaluation metrics have to be calculated only once and

not for each model seperately. Combining the biases and inter-model disagreement at different flow perecentiles allows almost similar conclusions as evaluating each model individually. Typically, the ensemble mean scatters around the true value (i.e. removing random noise by averaging) unless the error is systematic (Gudmundsson et al. 2012b). However, in this way if the error is not systematic, we lose track of the error to its specific model. In case the error is systematic the inter-model disagreement might diminish meaning that models agree on the error (e.g., Figs. A.11g, A.11n, and A.11u). In order to allow a process-based evaluation it is crucial that the ensemble members consider the same processes, albeit the mathematical representations may differ.

Since the existing 22 large-scale studies which evaluated the runoff estimations of multiple models (Table 1.1) often did not include the same models as well as different forcing, it is, hence, difficult to compare the results directly. Regarding this we call for the same suggestion as Beck et al. (2017a) and encourage efforts towards a single community hydrological model (Weiler and Beven 2015) for which it is possible to select alternative model structures. Beck et al. (2017a) proved in their study that the multi-parameterisation ensemble HBV-SIMREG (Beck et al. 2016) outperformed the multi-model ensemble. A single community hydrological model would facilitate the comparability of the results of different studies. Moreover, it would be unnecessary to set up, run, and maintain multiple models (Beck et al. 2017a). When focusing on a single community hydrological model, human resources, which have been used for one of the plethora of models, could be used instead to advance the community hydrological model.

5.2 Climatic and Physiographic Controls on Model Errors

The model evaluation already revealed "climatic" control of P biases in the forcing on the model errors. In this respect, we argue that the results of the statistical analysis for HLRs with complex topography should be interpreted very cautiously. Since model errors are corrupted by the forcing the statistical analysis would give answers for the wrong reasons. It is highly recommended to interpret results for which it is less likely that P errors propagate into the simulated runoff.

The regression analysis unveiled the following insights about the relationship between climatic and physiographic catchment characteristics and model errors:

- Subhumid plains/plateaus with permeable bedrock (HLR 8): The positive re-

lationship between $CORR$ and $B_{std-sqrt}$ Q_{95} (Fig. 4.11a) exhibits, that the simulated runoff tends to underestimate when there is a phase-shift (e.g., negative $CORR$) in the seasonality of supply and demand of water. This phase-shift suggests that rather an inadequate storage routine than the evapotranspiration routine might be responsible for that. Beck et al. (2013) found a negative relationship between $CORR$ and base flow index (BFI), which supports our hypothesis. SLO and $B_{std-sqrt}$ Q_{95} were negatively related (Fig. 4.11b) when a greater underestimation is present for steeper slopes. A steeper terrain and the geological setting of HLR 8 with a higher bedrock permeability suggest that lateral groundwater flow may be significant. Presumably, simulated runoff might be underestimated because steep sloping aquifers are drained too quickly. Yet, it is also likely that models do not account for lateral groundwater flow between grid-cells properly (Krakauer et al. 2014).

The relationship obtained between AI and B_{rel} Q_5 was postive where simulated runoff tends to be underestimated (overestimated) for low (high) AI values (Fig. 4.11c). The reason for this overestimation could be either that the models do not account for increasingly nonlinear response behaviour or that it is attributed to the positively biased P. The findings from Beck et al. (2017a) suggest that positive P errors are positively related to AI. Runoff underestimation in snow-influenced regions was already reported by Zaitchik et al. (2010). This is in agreement with the present results (Fig. 4.11d). The underestimation might result from shortcomings of the models when simulating the timing of snow accumulation and melt (Zaitchik et al. 2010). The relationship found between IRR and B_{rel} Q_5 seem to be heavily influenced by some outliers (Fig. 4.11e) and, thus, the interpretation may be misleading.

fLi_{ss} and B_{rel} Q_{75} were positively related (Fig. 4.11f). The simulated runoff appears to be underestimated (overestimated) for low (high) fLi_{ss}. Thus, simulated runoff tends to overestimate the Q_{75} for catchments with high fraction of fLi_{ss}. In general, siliciclastic sedimentary rocks represent, for example, sandstone, mudstone, and greywacke (Hartmann and Moosdorf 2012). Yet, since siliciclastic sedimentary rocks encompass both coarse grained and fine grained sediments the log-scale permeability may range from -12.5 (higher permeability) to -16.5 (lower permeability) (Gleeson et al. 2011). Due to this ambiguity it can only be speculated whether it is the lower or higher permeability that leads to the overestimation. Yet, since HLR 8 is characterised by permeable

bedrock the overestimation might originate from a higher permeable bedrock.

- (Sub-)Humid plains with very impermeable bedrock (HLR 9): The negative relationship found between P_{si} and $KGE_{\gamma\beta}$ (Fig. 4.11g) can be attributed to shortcomings of the models coping with greater seasonal precipitation dynamic.
- (Sub-)Humid plains/plateaus with impermeable bedrock (HLR 11): AI and $KGE_{\gamma\beta}$ were negatively related (Fig. 4.11h). This is in accord with Beck et al. (2017a) and Haddeland et al. (2011) which also observed that decrease of model performance is accompanied by increasing AI. This might be due to the interrelation between P biases in the forcing and $KGE_{\gamma\beta}$.

Certain Streamflow observations correspond to an area which is about one order of magnitude smaller than the corresponding grid cell or an area which is greater than 10000 km^2 (Fig. 2.4). In the latter case, channel routing effects might be present. Due to this scale mismatch model errors might potentially be controlled by the catchment area. However, the regression analysis between catchment size and model errors, proved that there is no systematic error. This is in line with Gudmundsson et al. (2012a).

Overall, among the informative predictor-predictand pairs, climate-related predictors were most favored. Additionally, they showed also the strongest relationships with the model errors (Figs. 4.9 and 4.11). Topography-related, land cover-related, and geology-related predictors appeared to be important only once, while predictors related to soils and water use were found to be relatively unimportant (Fig. 4.9). Generally, low R^2 values were obtained for bivariate relationships (Fig. 4.9), suggesting that multiple predictors have to be used to account adequately for the complex interplay between climatic and physiographic characteristics. Nonetheless, low R^2 values in Figure 4.9 were found either because the predictor under consideration has in fact no control or the quality of the dataset describing the predictor is not sufficient to make a relationship visible. In addition to the missing representation of the complex interplay between predictors, prevailing nonlinear relationships in Figure 4.11 corroborate the need for RF.

The obtained R^2 values of RF range from poor to moderate (Table A.2) and exceeded R^2 values of the regression analysis (Fig. 4.9). Despite those differences in R^2, relationships described by the partial dependence are in agreement with the relationships derived by the regression analysis (e.g., Figs. 4.12d and 4.11a).

However, results of RF are more complex than those of the regression analysis (Figs. 4.9 and 4.10). The patterns in the entire dataset on AI, $ELEV$, and $PERM$ additionally underpinned our strategy to partition the catchments into subsets based on these three dimensions. Through the RF approach the following climatic and physiographic controls on model errors are discernible:

- (Sub-)Humid plains with very permeable bedrock (HLR 3): Despite different controls found for $B_{std-sqrt}$ and B_{rel} (Fig. 4.10), $NDVI$ is highly relevant to both biases. Yet, presumably $NDVI$ is not a causal predictor for high flow biases. Due to its strong negative correlation with fS (Fig. A.1c) it acts as a surrogate and it is more likely fS to be causal. In an analogous manner, this might also hold for low flow biases. $NDVI$ correlates positively with PET and TA and negatively PET_{si}.

  The negative relationship derived from partial dependence on $NDVI$ for $\hat{B}_{std-sqrt}$ Q_5 (Fig. 4.12a) reflects the strong correlation between $NDVI$ and fS (Fig. A.1c). Hence, simulated runoff tends to overestimate Q_5 in snow-influenced catchments. Additionally, inter-model disagreement is controled by fS and increases for higher fS values (Fig. A.6h). This means that the error is not systematic (i.e., not all models perform poorly), but the averaging may not compensate for the increase of scattering. Regarding this, an explaination may be given by the differences in the snow routine (e.g., energy balance or degree day scheme, different number of layers; Table 2.1). The overestimation of Q_5 contradicts the findings of Gudmundsson et al. (2012a) and Zaitchik et al. (2010) which stated that simulated runoff is underestimated in snow-influenced regions. This contradiction is eventually due to the use of different metrics and/or the spatial disagreement between the catchment locations of the studies and HLR 3 (cf. Fig. 4.2). PET_{si} might also be noncausal since its strongly positively correlated with fS (Fig. A.1c). By contrast, $PERM$ and $SILT$ give evidence for causal relationships. $PERM$ is negatively related to $\hat{B}_{std-sqrt}$ Q_5 (Fig. 4.12a). Conversely, $SILT$ is positively related to $\hat{B}_{std-sqrt}$ Q_5 (Fig. 4.12a). Accordingly, simulated runoff has a tendency towards overestimation of Q_5 for less permeable bedrock and soils. It is conceivable that infiltration is erroneously parameterised and hence the models generate too much runoff.

  The interaction between $PERM$ and $NDVI$ extends from -13.0 to -12.6 ($PERM$) and 120 to 200 ($NDVI$) (Fig. 4.13a). Within this rectangular a

strong interaction occurs. Consequently, greatest overestimation occurs in combination with lower bedrock permeability and increasing influence of snow.

The positive relationship inferred from partial dependence on $NDVI$ for $\hat{B}_{std-sqrt}$ Q_{95} (Fig. 4.12b) manifests the strong correlation of $NDVI$ versus PET, $NDVI$ versus TA, and $NDVI$ versus PET_{si} (Fig. A.1c). This indicates that especially when energy is limited (low $NDVI$ and high fS) PET formulations might be responsible for the greater understimation. This was also reported by Beck et al. (2017a). Partial dependence on AI provides further support that the estimation of Q_{95} is worse when energy is limited (Fig. 4.12b). This error is more systematic since the inter-model disagreement is lower for lower AI (Figs. A.6i and A.2b). Another probable reason could be that too much surface runoff is generated and thus there is a lack of recharge. Partial dependence on $CLAY$ for $\hat{B}_{std-sqrt}$ Q_{95} unveiled a positive relationship, indicating that lower soil contents of clay amplify the underestimation (Fig. 4.12b). Since $CLAY$ and $SAND$ are strongly correlated (Fig. A.1c) the underestimation might be attributable to the soil storage routine (e.g., parameterisation of soil water capacity). Partial dependence on SLO for $\hat{B}_{std-sqrt}$ Q_{95} exposes an increase of underestimation for steeper slopes (Fig. 4.12b). A reason may be that models drain inclined aquifers too quickly.

The interaction between SLO and $NDVI$ restricts to the combination of lower SLO and higher $NDVI$ (Fig. 4.13b).

- Subhumid plains/plateaus with permeable bedrock (HLR 8): From the partial dependence on AI for $\hat{B}_{std-sqrt}$ Q_5 (Fig. 4.12c) a postive relationship was obtained. There are two potential explanations for this. First, models cannot cope with the nonlinear catchment response linked to an increasing AI. Second, positive P errors cause the overestimation. Concerning this, the negative relationship illustrated by the partial dependence on $ELEV$ (Fig. 4.12c) presents a counter argument. The positive $\hat{B}_{std-sqrt}$ Q_5 decreases for higher surface elevation. Partial dependence on PET_{si} shows almost no relationship. Furthermore, there is no physical reason for the positive relationship exhibited by the partial dependence on IRR for $\hat{B}_{std-sqrt}$ Q_5 (Fig. 4.12c).

  A reasonable interpretation of the interaction between AI and IRR is prevented since there is no physical meaning due to the relationship of IRR (Fig. 4.13c).

  The negative partial dependence on $ELEV$ and the positive partial depen-

dence on PET_{si} for $\hat{B}_{std-sqrt}$ Q_{95} (Fig. 4.12d) give evidence for underestimation when energy is limited. Both relations are considered to be noncausal. $ELEV$ might act as a proxy for energy availability although there is no correlation with TA apparent (Fig. A.1d). PET_{si} is strongly negatively correlated to PET (Fig. A.1d). A positive relationship was obtained from the partial dependence on $CORR$ and a negative relationship is described through the partial dependence on $PERM$ (Fig. 4.12d). These relationships let assume that models underestimate runoff release groundwater too quickly and/or evapotranspiration routine evaporates too much water. The latter case, may be even more distinct for energy-limited catchments (e.g., plateaus). Inter-model disagreement of Q_{95} is controled by AI (Fig. A.3d). Runoff simulations of the ten models diverge less for lower AI suggesting that the error is systematic. Moreover, fLi_{sc} has also control on inter-model disagreement. Runoff estimations agree less if there is a high fraction of carbonate sedimentary rocks. This means errors are less systematic.

The interaction between $ELEV$ and $CORR$ is clearly discernible (Fig. 4.13d). For $ELEV$ lower than 800 m MSL the two-variable partial dependence solely relies on $CORR$. Consequently, understimation is most pronounced if there are higher surface elevation and a phase-shifted seasonality of supply and demand of water.

- (Sub-)Humid plains with very impermeable bedrock (HLR 9): Partial dependence on $SILT$ for $\hat{B}_{std-sqrt}$ Q_{75} was found to be positive (Fig. 4.12e) indicating that soil storages of the models are oversized. This is reasonable since the catchments under consideration are characterised by a very low bedrock permeability highlighting the significance of soil storage. Here, interflow processes may be dominant. The slight negative partial dependence on AI (Fig. 4.12e) gives evidence that the overestimation may not be attributed to positive P errors in the forcing. Partial dependence on P_{si} levelled off around $\hat{B}_{std-sqrt}$ Q_{75} of 0 for values greater than 0.2 (Fig. 4.12e). For values less than 0.2 there, clearly, is a tendency towards overestimation. Surprisingly, models perform better if there is greater seasonality of precipitation. This was also found by partial dependence on P_{si} for $\hat{B}_{std-sqrt}$ Q_{95} (Fig. 4.12f). The diverging partial dependence of PET_{si} suggest that simulated runoff is more overestimated for values ranging from 0.4 to 0.6. However, PET_{si} is strongly correlated with PET and TA (Fig. A.1e). For both Q_{75} and Q_{95} inter-model disagreement is

controled by AI and P_{si} (Figs. A.2e and A.2e). Consequently, for lower values of AI and P_{si} errors are more systematic.

The two-variable partial dependence on P_{si} and $SILT$ illustrated that there is tendency to overestimation if precipitation seasonality is very low and percentage of silt soil content is relatively high (Fig. 4.13e). Conversely, there is a tendency for underestimation if precipitation seasonality is high and percentage of silt soil content is relatively low.

Positive relationships were derived by partial dependence on P and $SILT$ for $\hat{B}_{std-sqrt}$ Q_{95} (Fig. 4.12e). Apparently, the overestimation of the simulated runoff may be explained as such that soil storage is oversized and infiltration is overestimated.

The interaction between P_{si} and P was distinctly dominated by P_{si} (Fig. 4.13f). For values of P_{si} less than 0.2 $\hat{B}_{std-sqrt}$ Q_{95} is almost zero. Overestimation is most distinct for very low precipitation seasonality and P greater than 1000 $mm\ yr^{-1}$.

A surprising result is that ranks of permutation importances for $B_{std-sqrt}$ and B_{rel} 4.10 are not identical although the rank correlation between $B_{std-sqrt}$ and B_{rel} for the selected HLRs is greater than 0.95 (Table A.3). There are several possible reasons for these discrepancies. The mismatch is likely attributable to: (i) collinear predictor variables (Fig. A.1) and deficiencies of the RF algorithm to decorrelate them; (ii) distributions of B_{rel} are less bell-shaped than those of $B_{std-sqrt}$ (Fig. 4.8, Table 4.3); (iii) $B_{std-sqrt}$ and B_{rel} were differently standardised (see Sect. 3.3).

In spite of that fact, we came to the almost same conclusions when interpreting the partial dependencies of B_{rel} (Fig. A.2).

Overall, similar to the outcomes of the regression analysis the RF approach identified, too, climate-related predictors more frequently to be important. Nonetheless, the interplay with less frequent important predictors such as land cover, geology, and soils was essential to figure out how the model error is controlled by the climatic and physiographic characteristics. The links from evapotranspiration, geology and soils to the erroneous Q_{95} estimates for two HLRs (HLR 3 and HLR 8) give evidence that deficiencies in the evapotranspiration routine and storage routine of the state-of-the-art LHMs exist. Model errors found for Q_5 could be linked to a sensitivity for P errors and shortcomings in the snow routine. For one HLR (HLR 9) we demonstrated that errors have their origin in the soil routine. For all these reasons,

our recommendations to large-scale modellers are that they should examine these deficiencies and possibly make the necessary adjustments. Remedial actions could be: (i) models which do not implement a groundwater routine (e.g., JULES; Table 2.1) might add a routine to their model; (ii) existing groundwater routine have to be improved. The current approaches might not be sufficient to drain aquifers correctly. (iii) only 4 models (HBV-SIMREG, SWBM, LISFLOOD and WaterGAP3) incorporate calibration into the model setup. A calibration of the uncalibrated models may solve false parameterisation, though structural problems may persist. Beck et al. (2017a) proved that HBV-SIMREG, SWBM, LISFLOOD and WaterGAP3 benefited from the calibration procedure and outperformed their a priori parameters. Moreover, a priori parameters do not fulfill flux-matching (e.g., evapotranspiration) across spatial scales (Samaniego et al. 2017). A multiscale parameter regionalisation technique may therefore enhance model performance and reduce inter-model disagreement simultaneously (Samaniego et al. 2017).

5.3 Statistical Analysis: Critical Appraisal

The HLRs resulting from the K-means clustering showed that three simple indices $ELEV$, $PERM$, and AI, describing the three dimensions topography, geology, and climate, are good indicators for finding similar hydrologic landscapes. To further illustrate this, Figure 4.2 exhibits the spatial locations of the HLRs. Clearly, K-means clustering was capable of correct classification, for example, the Himalaya and the chalk catchments in southeastern UK (Fig. 4.2, Table 4.1). Despite the satisfactory partitioning of the catchments into subsets a few things might be criticised or improved. The three indices represent averaged catchment values and thus potential heterogeneity might be missed. Instead of using the catchment mean surface elevation, using the range of altitude may be a better indice to represent the catchment topography. This would have required further data assimilation which was not feasible within the given time period. One could also argue to use SLO instead of $ELEV$ but there was a strong correlation between these two variables (Fig. A.1a) and classification on $ELEV$ is easier to apply. Since the elbow and the number of selected clusters (Fig. 4.1) disagree, a sensitivity analysis could provide further insights on the stability of the analysis.

Notwithstanding a rather complex picture in terms of linkage between climatic and physiographic characteristics and model errors, the chosen methodology proved to be appropriate for evaluating the model error. However, regression analysis often

yielded too weak or no relationships (Fig. 4.9). This might be due to a high variance of the biases because positive and negative values are alternating. Analysing the variance of biases instead of analysing absolute biases was necessary to make the correct conclusions. Moreover, decomposing the biases into positive and negative biases was no alternative. The distributions would then have lost the bell-shape. This holds also true for absolute biases. Decomposed biases and absolute biases would have required transformation (e.g., logarithmic transformation). As a consequence, results would become less interpretable. Given the minor differences in the outcomes of the regression analysis and the RF analysis for HLR 8, both methods are recommendable. In general, an asset of the RF algorithm is that it accounts for interactions and nonlinearities among variables (Hastie et al. 2017). This allowed for a detection of complex patterns. Furthermore, RF is robust to noise and the bootstrap sampling reduces the uncertainty of the data (Bachmair et al. 2016; Hastie et al. 2017). Although we set the size of subsamples of predictors at each split m very small (see Sect. 3.5), RF could not fully decorrelate the predictors and showed a bias towards correlated variables (Strobl et al. 2008). In this respect, further model tuning (e.g., excluding collinear predictors, including only important predictors) might improve the RF performance. Generally, the obtained model accuracies of RF (Table A.2) might be more pessimistic than accuracies derived from cross-validation (Hastie et al. 2017). With the calculation of the partial dependence we could make the results of the RF analysis interpretable in a smiliar way as the regression analysis. A strenght of partial dependence was that interaction between important predictors could be illustrated. Nonetheless, the partly uneven shape of the partial dependence (e.g., non-monotony) was not always intuitive. Regarding this, a denser grid or a grid following the data density (e.g., individual grids for each decile) might be a remedy. Yet, this would further prolong the already high computation time.

5.4 Study Limitations

This study has several limitations, which can be improved in later research. First, we looked into model errors at a long-term. The results shown in this work form a good basis to delve for error patterns of simulated runoff in reproducing seasonal or inter-annual variability. Second, the results are limited to catchments with (sub-)humid climate. This is because of the lack of streamflow observations in (semi-)arid climates. Third, streamflow observations were not spatially uniform

available. Consequently, the evaluation was restricted to those catchments with available streamflow observations. Moreover, the (streamflow) datasets available for the model evaluation vary considerably in terms of accuracy and reliability (Sperna Weiland et al. 2015). Particularly, datasets do not provide any information on data quality. It may be worth striving for meta information describing data quality (e.g., distinction between poor and good quality). Fourth, average values for climatic and physiographic characteristics neglect potential heterogeneity. Future evaluation efforts should include this heterogeneity.

6 Conclusion

The main goal of this study was to unravel the link between climatic and physiographic characteristics and the origin of errors affecting large-scale hydrological models. By clustering the catchments into groups of hydrologic landscapes we identified three landscape settings for which errors could be ascribed to the model structure/parameterisation. Within these landscape settings the following controls and model deficiencies were found:

1. (Sub-)Humid plains with very permeable bedrock: The linkage of positive high flow biases and the fraction of snow cover pointed towards deficiencies in the snow routine while the linkage to geology and soils suggest deficiencies in the storage routine. Negatively biased low flow estimations correlated with the aridity index and the normalized difference vegetation index. Particularly, when energy is limited inadequacies in the evapotranspiration schemes were here found to be responsible.
2. Subhumid plains/plateaus with permeable bedrock: Due to the interrelation of high flow biases and aridity index P forcing errors could not be ruled out since aridity index P forcing errors may also be related. The snow routine was also found to be insufficient due to the negative correlation of high flow biases and the fraction of snow cover. Regarding the low flow biases, we identified two deficiencies. First, evapotranspiration routine behaves wrongly when energy is limited. Second, inadequacies in the storage routine cause errors when bedrock permeability is high and when seasonal correlation between water supply and demand is phase shifted.
3. (Sub-)humid plains with very impermeable bedrock: Overestimation of moderate low flow biases and low flow biases were likely induced by wrong parameterisation of soil storage and infiltration processes. Relations between soils and the biases indicate this.

Moreover, the clustering approach allowed us to reduce the likelihood that the error was caused by the forcing. Thereby, we identified the "climatic" control of biased P forcing on model errors for landscapes with complex topography while P biases

were less likely for landscapes with rather flat topography

We anticipate our study as a starting point for a more process-based model evaluation. In this respect, our approach might be employed on hyperresolution models and track the progress in model performance. Since we determined deficiencies in the snow, evapotranspiration, and storage routines a more tailored evaluation is necessary on this. Instead of solely focusing on long-term metrics, metrics on seasonal or inter-annual variability might provide useful insights.

References

Alcamo, J., Döll, P., Kaspar, F., and Siebert, S. (1997). *Global change and global scenarios of water use and availability: an application of WaterGAP 1.0*. Report. Center for Environmental Systems Research (CESR), University of Kassel, Germany.

Arnell, N. W. (1999). "A simple water balance model for the simulation of streamflow over a large geographic domain". In: *Journal of Hydrology* 217.3, pp. 314–335. DOI: 10.1016/S0022-1694(99)00023-2.

Bachmair, S., Svensson, C., Hannaford, J., Barker, L. J., and Stahl, K. (2016). "A quantitative analysis to objectively appraise drought indicators and model drought impacts". In: *Hydrol. Earth Syst. Sci.* 20.7, pp. 2589–2609. DOI: 10.5194/hess-20-2589-2016.

Barella-Ortiz, A., Polcher, J., Tuzet, A., and Laval, K. (2013). "Potential evaporation estimation through an unstressed surface-energy balance and its sensitivity to climate change". In: *Hydrol. Earth Syst. Sci.* 17.11, pp. 4625–4639. DOI: 10.5194/hess-17-4625-2013.

Beck, H. E., de Roo, A., and van Dijk, A. I. J. M. (2015). "Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments". In: *Journal of Hydrometeorology* 16.4, pp. 1478–1501. DOI: 10.1175/jhm-d-14-0155.1.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J. (2017a). "Global evaluation of runoff from 10 state-of-the-art hydrological models". In: *Hydrology and Earth System Sciences* 21.6, pp. 2881–2903. DOI: 10.5194/hess-21-2881-2017.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, A. L. (2016). "Global-scale regionalization of hydrologic model parameters". In: *Water Resources Research* 52.5, pp. 3599–3622. DOI: 10.1002/2015WR018247.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A. (2017b). "MSWEP: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data". In: *Hydrology and Earth System Sciences* 21.1, pp. 589–615. DOI: 10.5194/hess-21-589-2017.

Beck, H. E., van Dijk, A. I. J. M., Miralles, D. G., de Jeu, R. A. M., Bruijnzeel, L. A., McVicar, T. R., and Schellekens, J. (2013). "Global patterns in base flow index

and recession based on streamflow observations from 3394 catchments". In: *Water Resources Research* 49.12, pp. 7843–7863. DOI: 10.1002/2013WR013918.

Beven, K. J. (2011). *Rainfall-runoff modelling: the primer*. Chichester, England: John Wiley & Sons.

Bierkens, M. F. P. (2015). "Global hydrology 2015: State, trends, and directions". In: *Water Resources Research* 51.7, pp. 4923–4947. DOI: 10.1002/2015WR017173.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H. (2013). *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139235761.

Bontemps, S., Defourny, P., and van Bogaert, E. (2011). *GlobCover 2009 Products Description and Validation Report*. Report.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/a:1010933404324.

Brown, J., Ferrians, O. J., Heginbottom, J. A., and Melnikov, E. S. (1997). *Circum-Arctic Map of Permafrost and Ground-Ice Conditions*. [Available online at http://nsidc.org/data/ggd318]. Version 2. accessed 28 December 2012. National Snow and Ice Data Center.

Cloutier, V., Lefebvre, R., Therrien, R., and Savard, M. M. (2008). "Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system". In: *Journal of Hydrology* 353.3, pp. 294–313. DOI: 10.1016/j.jhydrol.2008.02.015.

Conover, W. (1971). *Practical nonparametric statistics*. New York: Wiley.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Marshall, C., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., and Meng, J. (2003). "Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project". In: *Journal of Geophysical Research: Atmospheres* 108.D22. DOI: 10.1029/2002JD003118.

Daly, C., Neilson, R. P., and Phillips, D. L. (1994). "A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain". In: *Journal of Applied Meteorology* 33.2, pp. 140–158. DOI: 10.1175/1520-0450(1994)033<0140:astmfm>2.0.co;2.

Decharme, B. (2007). "Influence of runoff parameterization on continental hydrology: Comparison between the Noah and the ISBA land surface models". In: *Journal of Geophysical Research: Atmospheres* 112.D19. DOI: doi:10.1029/2007JD008463.

Decharme, B. and Douville, H. (2006). "Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations". In: *Climate Dynamics* 27.7, pp. 695–713. DOI: 10.1007/s00382-006-0160-6.

Decharme, B. and Douville, H. (2007). "Global validation of the ISBA sub-grid hydrology". In: *Climate Dynamics* 29.1, pp. 21–37. DOI: 10.1007/s00382-006-0216-7.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F. (2011). "The ERA-Interim reanalysis: configuration and performance of the data assimilation system". In: *Quarterly Journal of the Royal Meteorological Society* 137.656, pp. 553–597. DOI: 10.1002/qj.828.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N. (2006). "GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface". In: *Bulletin of the American Meteorological Society* 87.10, pp. 1381–1398. DOI: 10.1175/bams-87-10-1381.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R. (2010). "GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States". In: *Ecology* 91.2, pp. 621–621. DOI: 10.1890/09-0889.1.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). "The Shuttle Radar Topography Mission". In: *Reviews of Geophysics* 45.2. DOI: 10.1029/2005RG000183.

Fowler, H. J. and Ekström, M. (2009). "Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes". In: *International Journal of Climatology* 29.3, pp. 385–416. DOI: 10.1002/joc.1827.

Gleeson, T., Smith, L., Moosdorf, N., Hartmann, J., Dürr, H. H., Manning, A. H., Beek, L. P. H. v., and Jellinek, A. M. (2011). "Mapping permeability over the surface of the Earth". In: *Geophysical Research Letters* 38.2. DOI: 10.1029/2010GL045565.

Gleeson, T., Wada, Y., Bierkens, M. F., and van Beek, L. P. (2012). "Water balance of global aquifers revealed by groundwater footprint". In: *Nature* 488.7410, pp. 197–200. DOI: 10.1038/nature11295.

Greuell, W., Andersson, J. C. M., Donnelly, C., Feyen, L., Gerten, D., Ludwig, F., Pisacane, G., Roudier, P., and Schaphoff, S. (2015). "Evaluation of five hydrological models across Europe and their suitability for making projections under climate change". In: *Hydrol. Earth Syst. Sci. Discuss.* 2015, pp. 10289–10330. DOI: 10.5194/hessd-12-10289-2015.

Gudmundsson, L. and Seneviratne, S. I. (2015). "Towards observation-based gridded runoff estimates for Europe". In: *Hydrol. Earth Syst. Sci.* 19.6, pp. 2859–2879. DOI: 10.5194/hess-19-2859-2015.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K. (2012a). "Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe". In: *Water Resources Research* 48.11. DOI: 10.1029/2011WR010911.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S. (2012b). "Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe". In: *Journal of Hydrometeorology* 13.2, pp. 604–620. DOI: 10.1175/jhm-d-11-083.1.

Guilyardi, E. (2006). "El Niño-mean state-seasonal cycle interactions in a multi-model ensemble". In: *Climate Dynamics* 26.4, pp. 329–348. DOI: 10.1007/s00382-005-0084-6.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *Journal of Hydrology* 377.1, pp. 80–91. DOI: 10.1016/j.jhydrol.2009.08.003.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P. (2011). "Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results". In: *Journal of Hydrometeorology* 12.5, pp. 869–884. DOI: 10.1175/2011jhm1324.1.

Hall, D. K., Salomonson, V. V., and Riggs, G. A. (2006). *MODIS/ Aqua snow cover daily L3 global 0.05deg CMG*. Version 5. accessed 11 November 2014. National Snow and Ice Data Center. DOI: 10.5067/EW53FPU9NAS6.

Hargreaves, G. L., Hargreaves, G. H., and Riley, J. P. (1985). "Irrigation Water Requirements for Senegal River Basin". In: *Journal of Irrigation and Drainage Engineering* 111.3, pp. 265–275. DOI: 10.1061/(ASCE)0733-9437(1985)111:3(265).

Harris, I., Jones, P., Osborn, T., and Lister, D. (2014). "Updated high-resolution grids of monthly climatic observations - the CRU-TS3.10 Dataset". In: *International Journal of Climatology* 34.3, pp. 623–642. DOI: 10.1002/joc.3711.

Hartmann, A., Gleeson, T., Rosolem, R., Pianosi, F., Wada, Y., and Wagener, T. (2015). "A large-scale simulation model to assess karstic groundwater recharge over Europe and the Mediterranean". In: *Geosci. Model Dev.* 8.6, pp. 1729–1746. DOI: 10.5194/gmd-8-1729-2015.

Hartmann, J. and Moosdorf, N. (2012). "The new global lithological map database GLiM: A representation of rock properties at the Earth surface". In: *Geochemistry, Geophysics, Geosystems* 13.12. DOI: 10.1029/2012GC004370.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition, corrected at 12th printing. Springer series in statistics. Springer, p. 745. DOI: 10.1007/978-0-387-84858-7.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G., and Gonzalez, M. R. (2014). "SoilGrids1km - Global Soil Information Based on Automated Mapping". In: *PLOS ONE* 9.8. DOI: 10.1371/journal.pone.0105992.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). "Very high resolution interpolated climate surfaces for global land areas". In: *International Journal of Climatology* 25.15, pp. 1965–1978. DOI: 10.1002/joc.1276.

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S. (2013). "Global flood risk under climate change". In: *Nature Climate Change* 3, p. 816. DOI: 10.1038/nclimate1911.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*. Corrected at 8th printing. Springer texts in statistics. Springer, p. 426. DOI: 10.1007/978-1-4614-7138-7.

Kling, H., Fuchs, M., and Paulin, M. (2012). "Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios". In: *Journal of Hydrology* 424-425, pp. 264–277. DOI: 10.1016/j.jhydrol.2012.01.011.

Knoben, W. J. M., Woods, R. A., and Freer, J. E. (2018). "A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data". In: *Water Resources Research* 57.7. DOI: 10.1029/2018WR022913.

Krakauer, N. Y., Haibin, L., and Ying, F. (2014). "Groundwater flow across spatial scales: importance for climate modeling". In: *Environmental Research Letters* 9.3, p. 034003. DOI: 10.1088/1748-9326/9/3/034003.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S. (2000). "Multimodel Ensemble Forecasts for Weather and Seasonal Climate". In: *Journal of Climate* 13.23, pp. 4196–4216. DOI: 10.1175/1520-0442(2000)013<4196:meffwa>2.0.co;2.

Lehner, B. and Döll, P. (2004). "Development and validation of a global database of lakes, reservoirs and wetlands". In: *Journal of Hydrology* 296.1, pp. 1–22. DOI: 10.1016/j.jhydrol.2004.03.028.

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Cosgrove, B. A., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D. (2004). "Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project". In: *Journal of Geophysical Research: Atmospheres* 109.D7. DOI: 10.1029/2003JD003517.

Materia, S., Dirmeyer, P. A., Guo, Z., Alessandri, A., and Navarra, A. (2010). "The Sensitivity of Simulated River Discharge to Land Surface Representation and Meteorological Forcings". In: *Journal of Hydrometeorology* 11.2, pp. 334–351. DOI: 10.1175/2009jhm1162.1.

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., Uijlenhoet, R., and Teuling, A. J. (2018). "Mapping (dis)agreement in hydrologic projections". In: *Hydrol. Earth Syst. Sci.* 22.3, pp. 1775–1791. DOI: 10.5194/hess-22-1775-2018.

Milly, P. C. D., Dunne, K. A., and Vecchia, A. V. (2005). "Global pattern of trends in streamflow and water availability in a changing climate". In: *Nature* 438, p. 347. DOI: 10.1038/nature04312.

Nash, J. E. and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models part I - A discussion of principles". In: *Journal of Hydrology* 10.3, pp. 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Pappenberger, F., Dutra, E., Wetterhall, F., and Cloke, H. L. (2012). "Deriving global flood hazard maps of fluvial floods through a physical model cascade". In: *Hydrol. Earth Syst. Sci.* 16.11, pp. 4143–4156. DOI: 10.5194/hess-16-4143-2012.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peel, M. C., Chiew, F. H., Western, A. W., and McMahon, T. A. (2000). *Extension of unimpaired monthly streamflow data and regionalisation of parameter values to*

estimate streamflow in ungauged catchments. Report. Cent. for Environ. Appl. Hydrol., Univ. of Melbourne.

Petersen, T., Devineni, N., and Sankarasubramanian, A. (2012). "Seasonality of monthly runoff over the continental United States: Causality and relations to mean annual and mean monthly distributions of moisture and energy". In: *Journal of Hydrology* 468-469, pp. 139–150. DOI: 10.1016/j.jhydrol.2012.08.028.

Pokhrel, Y. N., Hanasaki, N., Yeh, P. J. F., Yamada, T. J., Kanae, S., and Oki, T. (2013). "Model estimates of sea-level change due to anthropogenic impacts on terrestrial water storage". In: *Nature Geoscience* 5, p. 389. DOI: 10.1038/ngeo1476.

Prudhomme, C., Parry, S., Hannaford, J., Clark, D. B., Hagemann, S., and Voss, F. (2011). "How Well Do Large-Scale Models Reproduce Regional Hydrological Extremes in Europe?" In: *Journal of Hydrometeorology* 12.6, pp. 1181–1204. DOI: 10.1175/2011jhm1387.1.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D. (2004). "The Global Land Data Assimilation System". In: *Bulletin of the American Meteorological Society* 85.3, pp. 381–394. DOI: 10.1175/bams-85-3-381.

Rogers, E., Parris, D., and DiMego, G. (1999). *Changes to the NCEP operational Eta Analysis, technical procedures bulletin*. Report. Off. of Meteorol., Natl. Weather Serv.

Rust, H. W., Kruschke, T., Dobler, A., Fischer, M., and Ulbrich, U. (2015). "Discontinuous Daily Temperatures in the WATCH Forcing Datasets". In: *Journal of Hydrometeorology* 16.1, pp. 465–472. DOI: 10.1175/jhm-d-14-0123.1.

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E., and Warrach-Sagi, K. (2017). "Toward seamless hydrologic predictions across spatial scales". In: *Hydrology and Earth System Sciences* 21.9, pp. 4323–4346. DOI: 10.5194/hess-21-4323-2017.

Schaefli, B. and Gupta, H. V. (2007). "Do Nash values have value?" In: *Hydrological Processes* 21.15, pp. 2075–2080. DOI: 10.1002/hyp.6825.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Weiland, F. S., Minvielle, M., Calvet, J.-C., Decharme, B., and Eisner, S. (2017). "A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset". In: *Earth System Science Data* 9.2, pp. 389–413. DOI: 10.5194/essd-9-389-2017.

Schellnhuber, H. J., Frieler, K., and Kabat, P. (2014). "The elephant, the blind, and the intersectoral intercomparison of climate impacts". In: *Proceedings of*

the National Academy of Sciences 111.9, pp. 3225–3227. DOI: 10.1073/pnas.1321791111.

Sheffield, J., Goteti, G., and Wood, E. F. (2006). "Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling". In: *Journal of Climate* 19.13, pp. 3088–3111. DOI: 10.1175/jcli3790.1.

Siebert, S., Henrich, V., Frenken, K., and Burke, J. (2013). *Global Map of Irrigation Areas version 5*. Report. Rheinische Friedrich-Wilhelms-University / Food and Agriculture Organization of the United Nations.

Sperna Weiland, F. C., Vrugt, J. A., van Beek, R. P. H., Weerts, A. H., and Bierkens, M. F. P. (2015). "Significant uncertainty in global scale hydrological modeling from precipitation data errors". In: *Journal of Hydrology* 529, pp. 1095–1115. DOI: 10.1016/j.jhydrol.2015.08.061.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1, p. 307. DOI: 10.1186/1471-2105-9-307.

Tallaksen, L. M. and Stahl, K. (2014). "Spatial and temporal patterns of large-scale droughts in Europe: Model dispersion and performance". In: *Geophysical Research Letters* 41.2, pp. 429–434. DOI: 10.1002/2013GL058573.

Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E. (2013). "Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide". In: *Water Resources Research* 49.5, pp. 2729–2746. DOI: 10.1002/wrcr.20251.

Van Vliet, M. T. H., van Beek, L. P. H., Eisner, S., Flörke, M., Wada, Y., and Bierkens, M. F. P. (2016). "Multi-model assessment of global hydropower and cooling water discharge potential under climate change". In: *Global Environmental Change* 40, pp. 156–170. DOI: 10.1016/j.gloenvcha.2016.07.007.

Vörösmarty, C. J., Federer, C. A., and Schloss, A. L. (1998). "Potential evaporation functions compared on US watersheds: Possible implications for global-scale water balance and terrestrial ecosystem modeling". In: *Journal of Hydrology* 207.3, pp. 147–169. DOI: 10.1016/S0022-1694(98)00109-7.

Wada, Y., Beek, L. P. H., Weiland, F. C. S., Chao, B. F., Wu, Y.-H., and Bierkens, M. F. P. (2012). "Past and future contribution of global groundwater depletion to sea-level rise". In: *Geophysical Research Letters* 39.9. DOI: 10.1029/2012GL051230.

Wada, Y., van Beek, L. P., van Kempen, C. M., Reckman, J. W., Vasak, S., and Bierkens, M. F. (2010). "Global depletion of groundwater resources". In: *Geophysical Research Letters* 37.20. DOI: 10.1029/2010GL044571.

Walsh, R. P. D. and Lawler, D. M. (1981). "Rainfall Seasonality: Description, Spatial Patterns And Change Through Time". In: *Weather* 36.7, pp. 201–208. DOI: 10.1002/j.1477-8696.1981.tb05400.x.

Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., van Beek, R., Bierkens, M. F. P., Ligtvoet, W., and Winsemius, H. C. (2013). "Assessing flood risk at the global scale: model setup, results, and sensitivity". In: *Environ. Res. Lett.* 8.4, p. 044019. DOI: 10.1088/1748-9326/8/4/044019.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M. (2011). "Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century". In: *Journal of Hydrometeorology* 12.5, pp. 823–848. DOI: 10.1175/2011JHM1369.1.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P. (2014). "The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data". In: *Water Resources Research* 50.9, pp. 7505–7514. DOI: 10.1002/2014WR015638.

Weiler, M. and Beven, K. (2015). "Do we need a Community Hydrological Model?" In: *Water Resources Research* 51.9, pp. 7777–7784. DOI: 10.1002/2014WR016731.

Winter, T. C. (2001). "The Concept of Hydrologic Landscapes". In: *JAWRA Journal of the American Water Resources Association* 37.2, pp. 335–349. DOI: 10.1111/j.1752-1688.2001.tb00973.x.

Wolock, D. M., Winter, T. C., and McMahon, G. (2004). "Delineation and Evaluation of Hydrologic-Landscape Regions in the United States Using Geographic Information System Tools and Multivariate Statistical Analyses". In: *Environmental Management* 34.1, pp. 71–88. DOI: 10.1007/s00267-003-5077-9.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D. (2012). "Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow". In: *Journal of Geophysical Research: Atmospheres* 117.D3. DOI: 10.1029/2011JD016051.

Yang, H., Piao, S., Zeng, Z., Ciais, P., Yin, Y., Friedlingstein, P., Sitch, S., Ahlström, A., Guimberteau, M., Huntingford, C., Levis, S., Levy, P. E., Huang, M., Li, Y., Li, X., Lomas, M. R., Peylin, P., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Zhao, F., and Wang, L. (2015). "Multicriteria evaluation of discharge simulation in Dynamic Global Vegetation Models". In: *Journal of Geophysical Research: Atmospheres* 120.15, pp. 7488–7505. DOI: 10.1002/2015JD023129.

Yu, Y., Disse, M., Yu, R., Yu, G., Sun, L., Huttner, P., and Rumbaur, C. (2015). "Large-Scale Hydrological Modeling and Decision-Making for Agricultural Wa-

ter Consumption and Allocation in the Main Stem Tarim River, China". In: *Water* 7.6, p. 2821. DOI: 10.3390/w7062821.

Zaitchik, B. F., Rodell, M., and Olivera, F. (2010). "Evaluation of the Global Land Data Assimilation System using global river discharge data and a source-to-sink routing scheme". In: *Water Resources Research* 46.6. DOI: 10.1029/2009WR007811.

Zhang, Y., Zheng, H., Chiew, F. H. S., Peña-Arancibia, J., and Zhou, X. (2016). "Evaluating Regional and Global Hydrological Models against Streamflow and Evapotranspiration Measurements". In: *Journal of Hydrometeorology* 17.3, pp. 995–1010. DOI: 10.1175/jhm-d-15-0107.1.

Zhou, X., Zhang, Y., Wang, Y., Zhang, H., Vaze, J., Zhang, L., Yang, Y., and Zhou, Y. (2012). "Benchmarking global land surface models against the observed mean annual runoff from 150 large basins". In: *Journal of Hydrology* 470-471, pp. 269–279. DOI: 10.1016/j.jhydrol.2012.09.002.

A Appendix

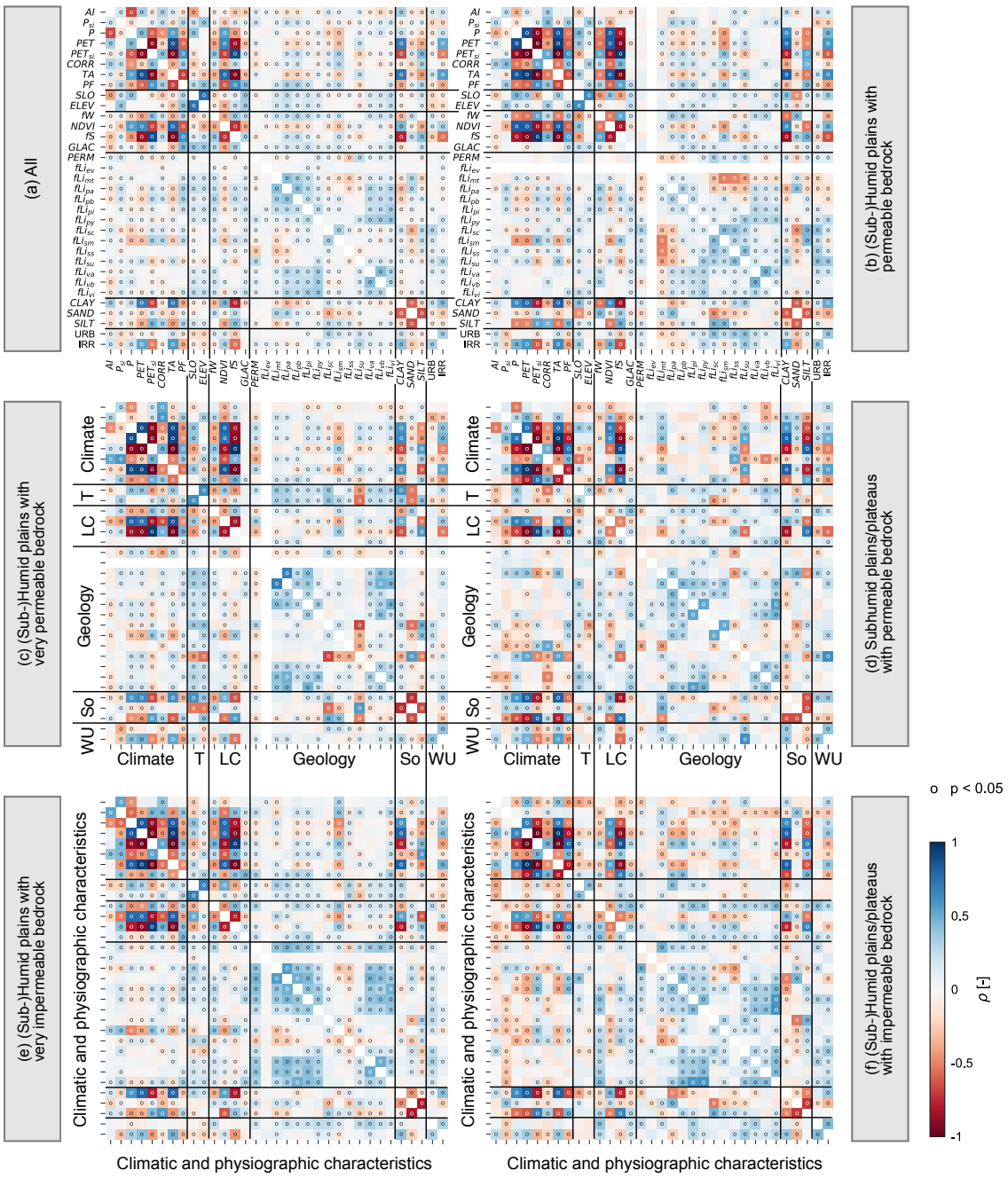


Figure A.1: Rank correlation coefficients (ρ) of climatic and physiographic characteristics for the entire dataset set (a) and for selected HLRs (b-f). Blue (red) indicates a positive correlation (negative correlation). *o* displays significant correlation at the 5% level.

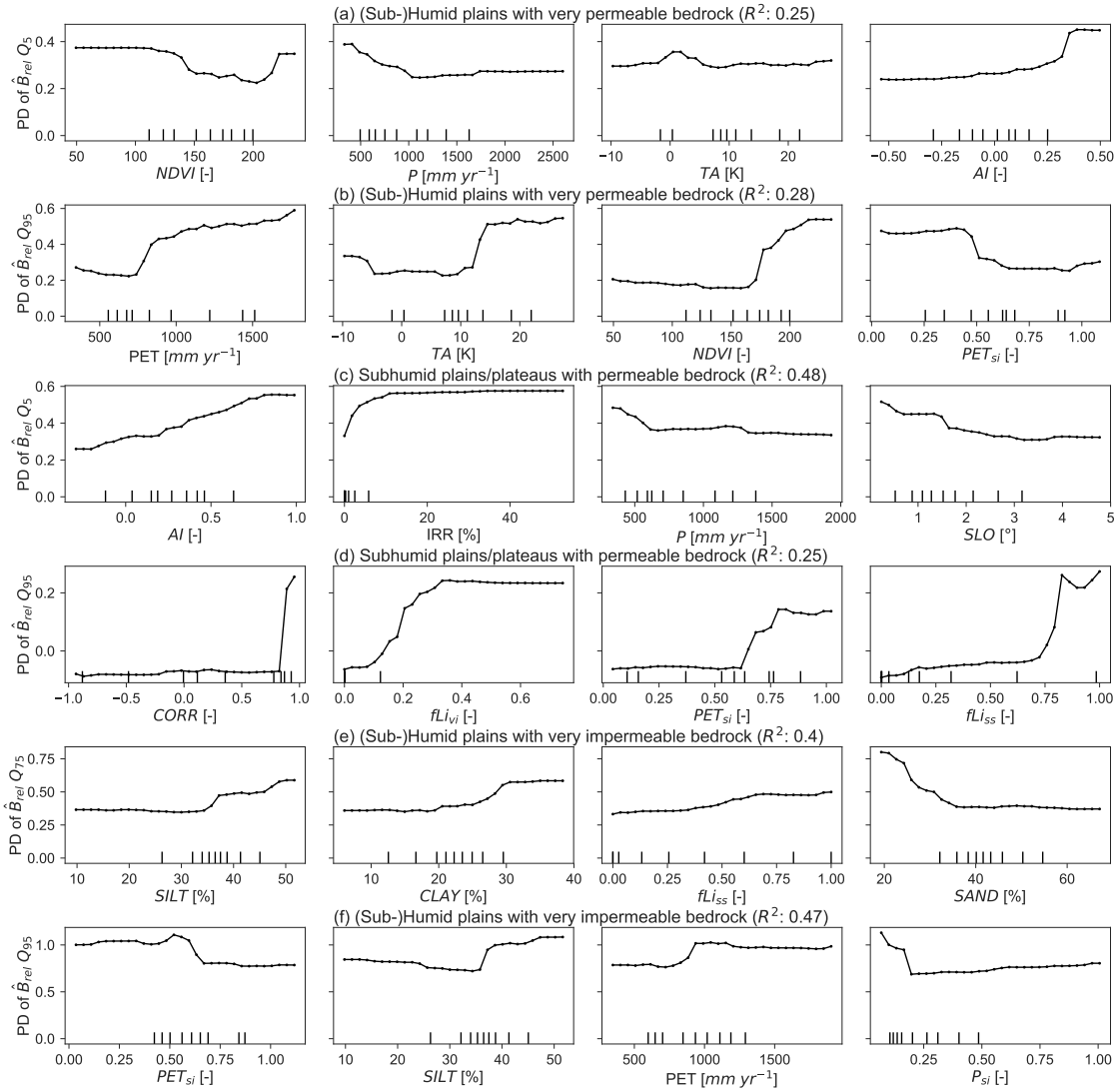


Figure A.2: Single-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis) versus partial dependence of $\hat{B}_{std-sqrt}$ Q_5, $\hat{B}_{std-sqrt}$ Q_{75}, and $\hat{B}_{std-sqrt}$ Q_{95} (along the y-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). The hash marks at the base of the plots delineate deciles of the corresponding predictor variable. R^2 exhibits the out-of-bag accuracy. The hat ($\hat{}$) denotes the predicted metric by the RF. For abbreviations on x-axis and y-axis see Table 2.2.

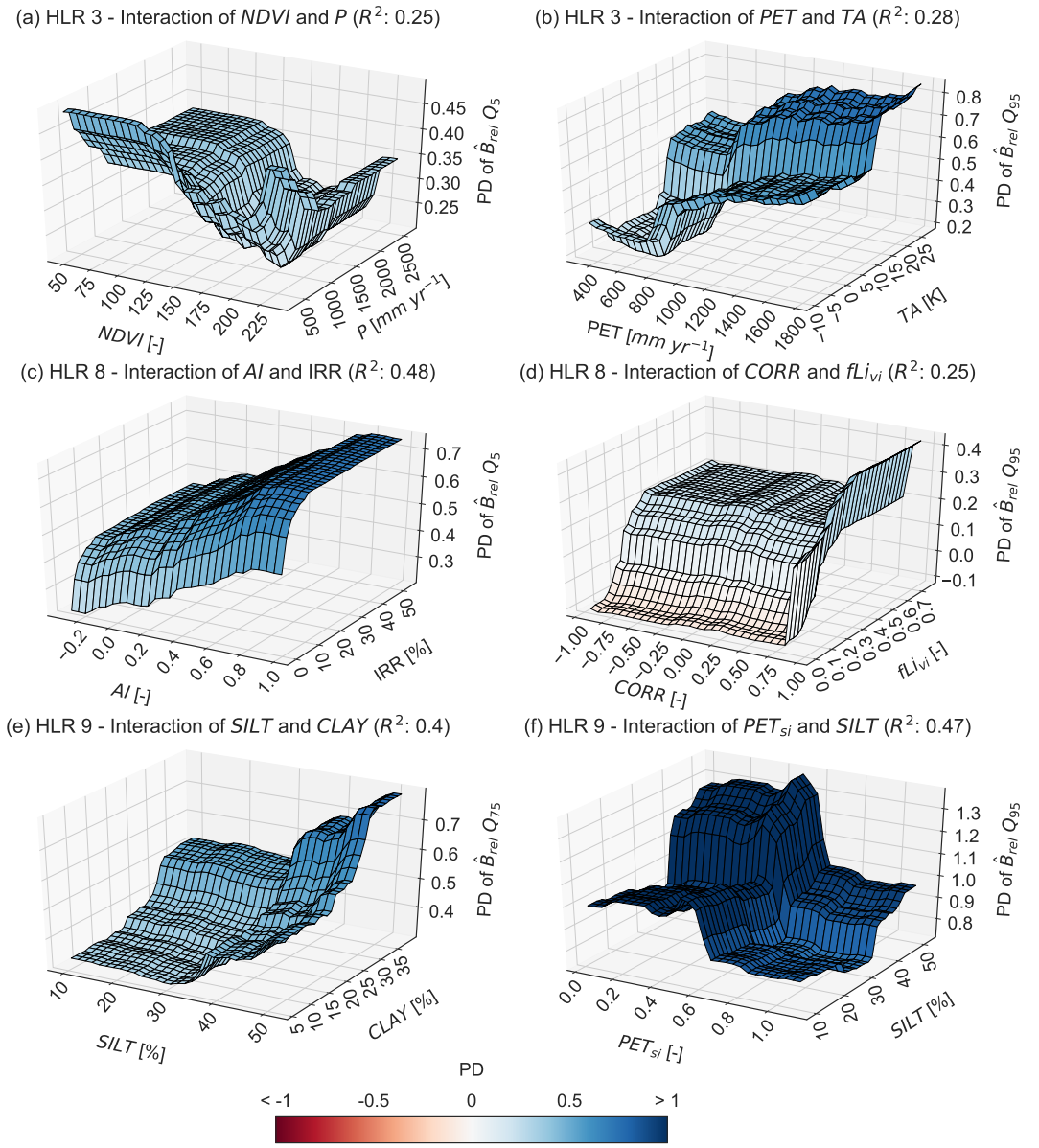


Figure A.3: Two-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis and y-axis) versus partial dependence of $\hat{B}_{std-sqrt}$ Q_5, $\hat{B}_{std-sqrt}$ Q_{75}, and $\hat{B}_{std-sqrt}$ Q_{95} (along the z-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). Predictors are sorted descendingly by ranks of importance from left to right. R^2 exhibits the out-of-bag accuracy. The hat (ˆ) denotes the predicted metric by the RF. For abbreviations on x-axis and y-axis see Table 2.2 and for descriptions of HLRs see Table 4.1.

Table A.1: Coefficients of determination (R^2) of bivariate regression between evaluation metrics and catchment size

Metric	All	HLR 1	HLR 2	HLR 3	HLR 4	HLR 5	HLR 6	HLR 7	HLR 8	HLR 9	HLR 10	HLR 11	HLR 12
$B_{std-sqrt}$ Q_5	0.02	0.16	0.01	0.03	0.02	0.09	0.03	0.25	0.09	0.01	0.13	0.03	0.11
$B_{std-sqrt}$ Q_{25}	0	0.04	0	0.01	0.01	0	0.01	0.19	0.08	0	0.08	0.04	0.03
$B_{std-sqrt}$ Q_{50}	0	0.01	0	0.02	0.01	0	0	0.01	0.05	0.01	0	0.02	0.01
$B_{std-sqrt}$ Q_{75}	0	0	0.02	0.01	0	0.01	0.01	0.13	0.03	0.03	0.04	0.01	0.01
$B_{std-sqrt}$ Q_{95}	0.01	0.01	0.04	0.01	0	0.03	0	0.15	0.05	0.06	0.04	0	0.02
B_{rel} Q_5	0.04	0.01	0.01	0.02	0.04	0.1	0.04	0.2	0.08	0.02	0.08	0.03	0.01
B_{rel} Q_{25}	0.02	0.01	0.01	0	0.01	0.02	0.02	0.33	0.04	0.01	0.02	0.08	0.01
B_{rel} Q_{50}	0	0.03	0	0.01	0	0.01	0	0	0.02	0.01	0	0.06	0.02
B_{rel} Q_{75}	0.01	0.05	0.02	0.02	0.01	0.03	0	0.06	0	0.02	0.03	0.02	0.01
B_{rel} Q_{95}	0.02	0.05	0.05	0.01	0.02	0.05	0.01	0.02	0	0.06	0.03	0	0.02
$KGE_{\gamma\beta}$	0	0.01	0	0	0.01	0	0	0.3	0.01	0.02	0.17	0.03	0.03
CV Q_5	0.02	0.01	0	0.1	0.03	0.07	0.06	0.09	0.01	0.03	0.04	0	0.01
CV Q_{25}	0.01	0.01	0	0.04	0.02	0.03	0.02	0	0	0.01	0	0	0.02
CV Q_{50}	0.01	0.01	0.01	0.07	0.01	0	0.02	0.02	0	0.01	0.14	0	0
CV Q_{75}	0	0.06	0	0	0	0.01	0.03	0.23	0	0.01	0.05	0.02	0.05
CV Q_{95}	0	0.02	0	0	0.01	0.01	0.01	0	0.01	0	0.02	0.04	0.01

Table A.2: Out-of-bag accuracy (R^2) of random forest for entire data and HLRs for all evaluation metrics

Metric	All	HLR 1	HLR 2	HLR 3	HLR 4	HLR 5	HLR 6	HLR 7	HLR 8	HLR 9	HLR 10	HLR 11	HLR 12
$B_{std-sqrt}$ Q_5	0.44	0.38	0.16	0.22	0.47	0.38	0.17	0.47	0.35	0.25	0.43	0.06	0.19
$B_{std-sqrt}$ Q_{25}	0.35	0.19	0.03	0.18	0.38	0.22	0.06	0.28	0.23	0.34	0.35	0.03	0.4
$B_{std-sqrt}$ Q_{50}	0.28	0.2	0.04	0.15	0.14	0.32	0.09	0.07	0.25	0.34	0.48	0.04	0.38
$B_{std-sqrt}$ Q_{75}	0.44	0.35	0.28	0.18	0.4	0.47	0.11	0.27	0.34	0.42	0.54	0.12	0.45
$B_{std-sqrt}$ Q_{95}	0.51	0.41	0.39	0.28	0.5	0.49	0.11	0.33	0.33	0.54	0.55	0.2	0.43
B_{rel} Q_5	0.46	0.31	0.3	0.25	0.49	0.38	0.06	0.36	0.48	0.39	0.43	0.4	0.2
B_{rel} Q_{25}	0.39	0.11	0.29	0.31	0.35	0.18	0.02	0.38	0.23	0.44	0.37	0.34	0.33
B_{rel} Q_{50}	0.33	0.07	0.24	0.18	0.24	0.25	0.16	0.17	0.2	0.33	0.35	0.39	0.33
B_{rel} Q_{75}	0.4	0.28	0.25	0.24	0.35	0.44	0.23	0.2	0.28	0.4	0.45	0.33	0.31
B_{rel} Q_{95}	0.42	0.24	0.36	0.28	0.37	0.43	0.09	0.24	0.25	0.47	0.35	0.12	0.26
$KGE_{\gamma\beta}$	0.46	0.24	0.37	0.32	0.37	0.45	-0.14	0.4	0.18	0.48	0.18	0.39	0.21
CV Q_5	0.45	0.31	0.4	0.48	0.44	0.51	-0.08	0.21	0.33	0.37	0.27	0.26	0.21
CV Q_{25}	0.46	0.21	0.24	0.43	0.3	0.25	-0.19	0.29	0.46	0.38	0.24	0.34	0.22
CV Q_{50}	0.56	0.2	0.29	0.39	0.35	0.21	0.17	0.16	0.35	0.49	0.15	0.49	0.28
CV Q_{75}	0.65	0.41	0.33	0.45	0.43	0.42	0.23	0.31	0.28	0.5	0.48	0.6	0.3
CV Q_{95}	0.69	0.55	0.56	0.53	0.55	0.51	0.27	0.28	0.46	0.59	0.45	0.62	0.4

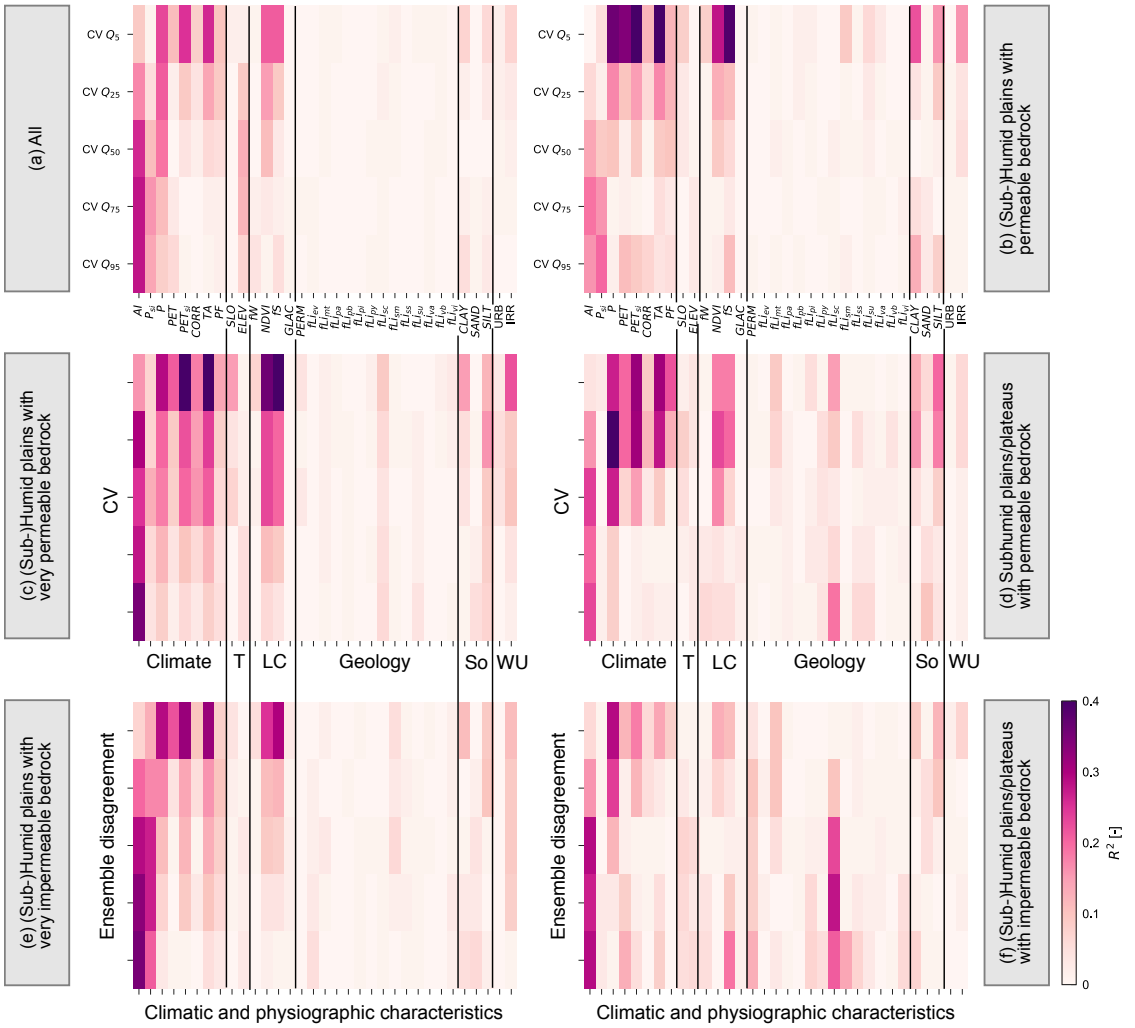


Figure A.4: Coefficients of determination (R^2) of bivariate (non-)linear regression for CV. Heatmaps of the R^2 are depicted for the entire dataset (a) and for selected HLRs (b-f). Purple (white) indicates fair relationship (no relationship). Abbreviations referring to: T, Topography; So, Soils; WU, Water use.

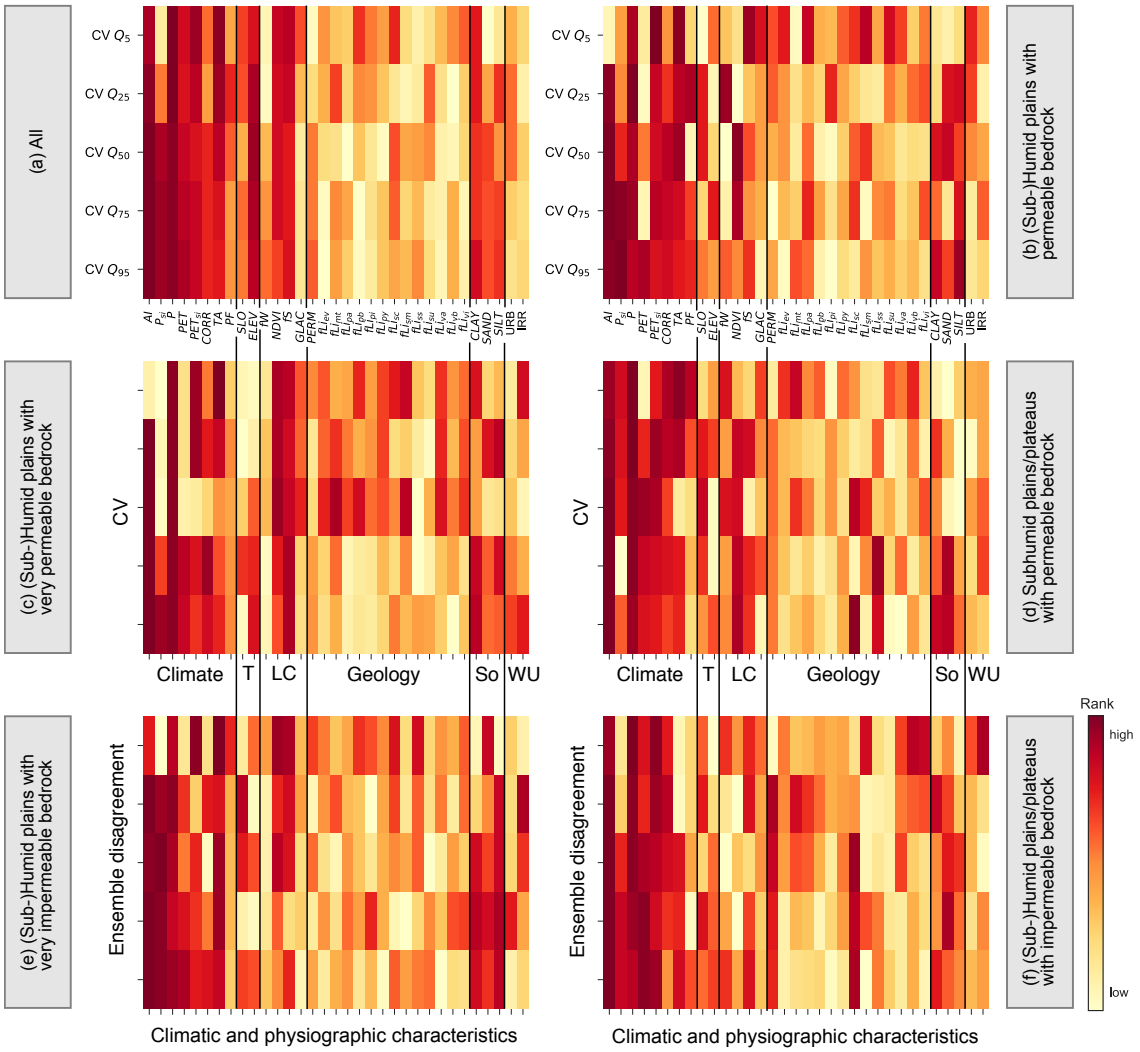


Figure A.5: Ranks of permutation importance of random forest for CV. Heatmaps of the ranks are depicted for the entire dataset (a) and for selected HLRs (b-f). Red (yellow) displays high (low) ranks. Abbreviations referring to: T, Topography; So, Soils; WU, Water use.

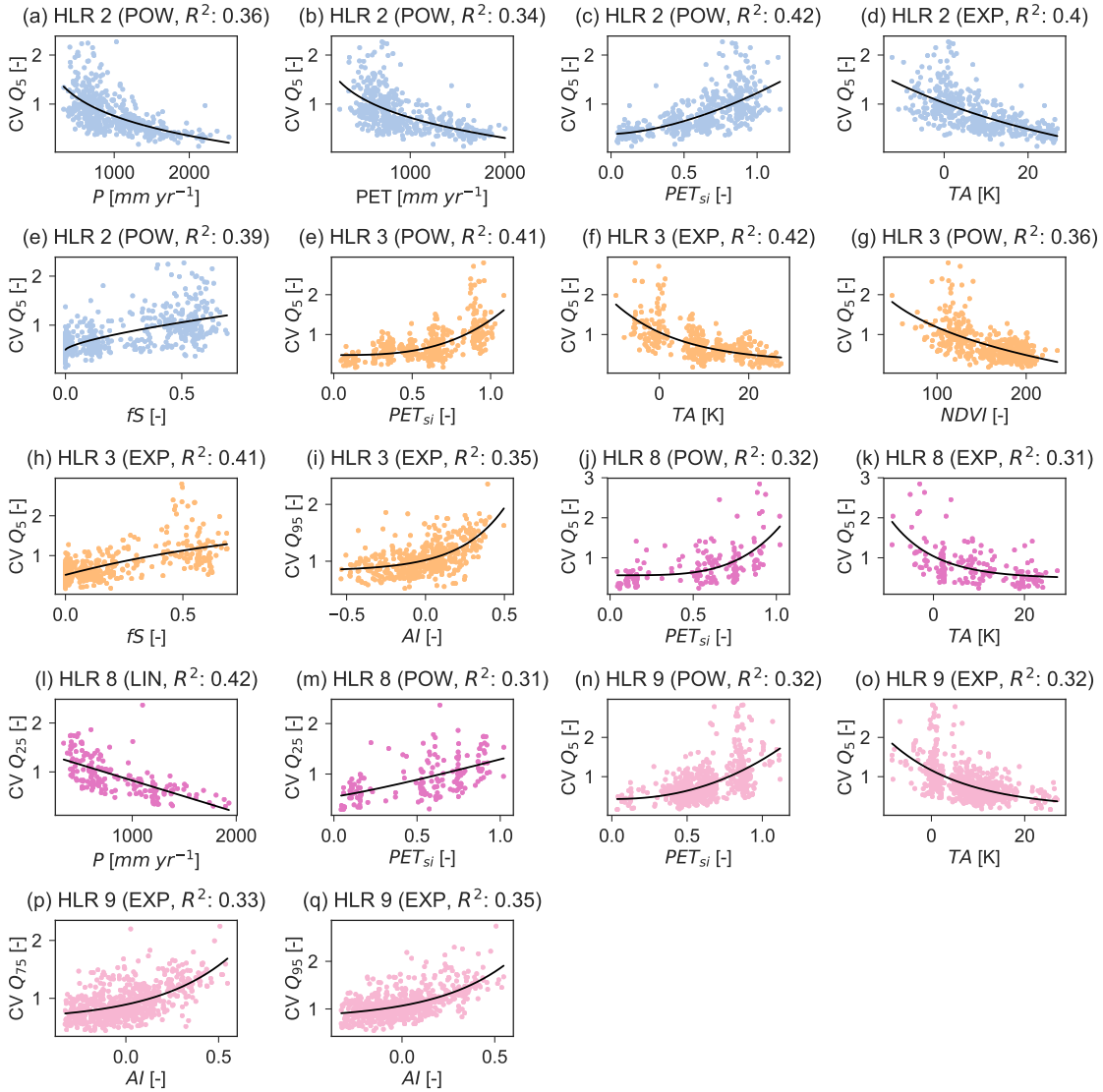


Figure A.6: Scatterplots of climatic and physiographic characteristics (along the x-axis) versus CV (along the y-axis), including the best-fit regression. Scatterplots are shown for $R^2 > 0.3$. Each data point represents a catchment. Abbreviations referring to the HLR (for description see Table 4.3) and to the type of the best-fit regression function: EXP, exponential; LIN, linear; LOG, logarithmic; and POW, power.

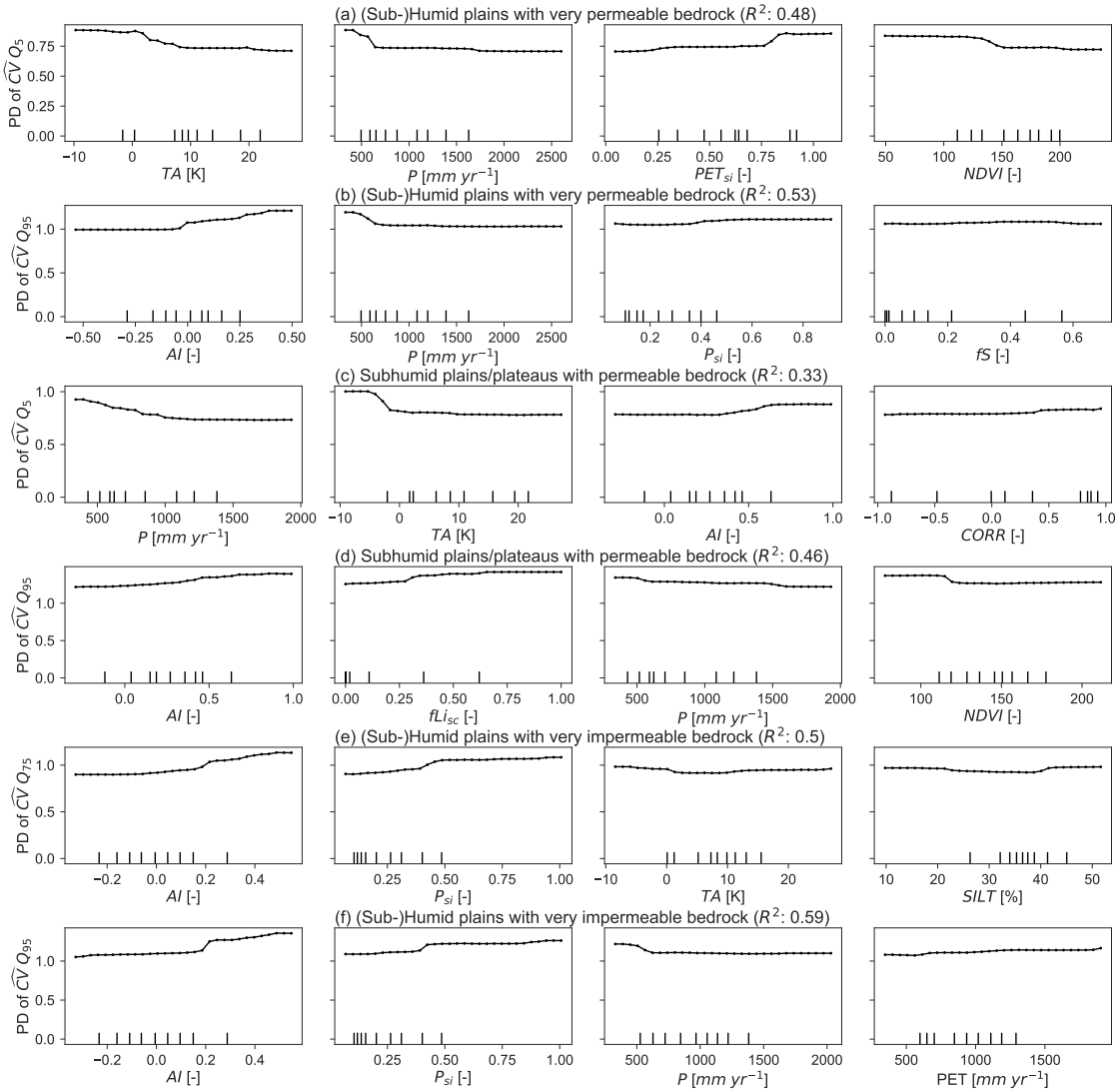


Figure A.7: Single-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis) versus partial dependence of $\hat{CV}\ Q_5$, $\hat{CV}\ Q_{75}$, and $\hat{CV}\ Q_{95}$ (along the y-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). The hash marks at the base of the plots delineate deciles of the corresponding predictor variable. R^2 exhibits the out-of-bag accuracy. The hat ($\hat{}$) denotes the predicted metric by the RF. For abbreviations on x-axis and y-axis see Table 2.2.

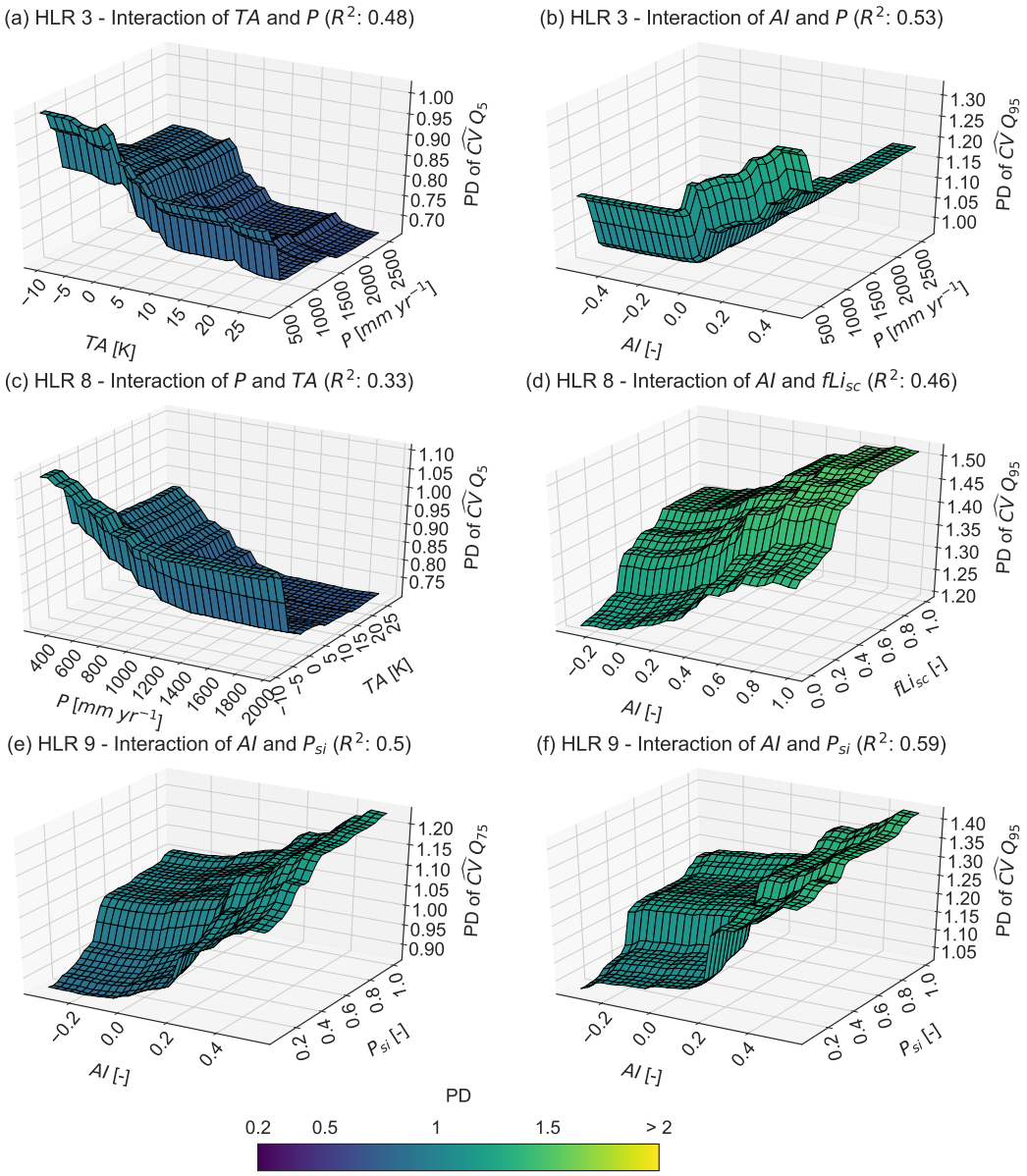


Figure A.8: Two-variable partial dependence (PD) plots of climatic and physiographic characteristics (along the x-axis and y-axis) versus partial dependence of $\hat{CV}\ Q_5$, $\hat{CV}\ Q_{75}$, and $\hat{CV}\ Q_{95}$ (along the z-axis), respectively. Plots are shown for HLR 3 (a, b), HLR 8 (c, d), and HLR 9 (e, f). Predictors are sorted descendingly by ranks of importance from left to right. R^2 exhibits the out-of-bag accuracy. The hat ($\hat{}$) denotes the predicted metric by the RF. For abbreviations on x-axis and y-axis see Table 2.2 and for descriptions of HLRs see Table 4.1.

Table A.3: Rank correlation between $B_{std-sqrt}$ and B_{rel} for all five flow percentiles in the entire dataset and HLRs

	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}
All	0.98	0.97	0.97	0.96	0.96
HLR 1	0.99	0.98	0.97	0.98	0.97
HLR 2	0.98	0.99	0.98	0.98	0.98
HLR 3	0.98	0.97	0.98	0.99	0.98
HLR 4	0.98	0.98	0.97	0.97	0.96
HLR 5	0.99	0.99	0.99	0.98	0.98
HLR 6	0.98	0.98	0.98	0.98	0.98
HLR 7	0.98	0.98	0.97	0.96	0.95
HLR 8	0.95	0.96	0.98	0.97	0.96
HLR 9	0.99	0.98	0.96	0.97	0.97
HLR 10	0.95	0.91	0.93	0.9	0.89
HLR 11	0.97	0.97	0.97	0.96	0.95
HLR 12	0.97	0.99	0.98	0.98	0.98

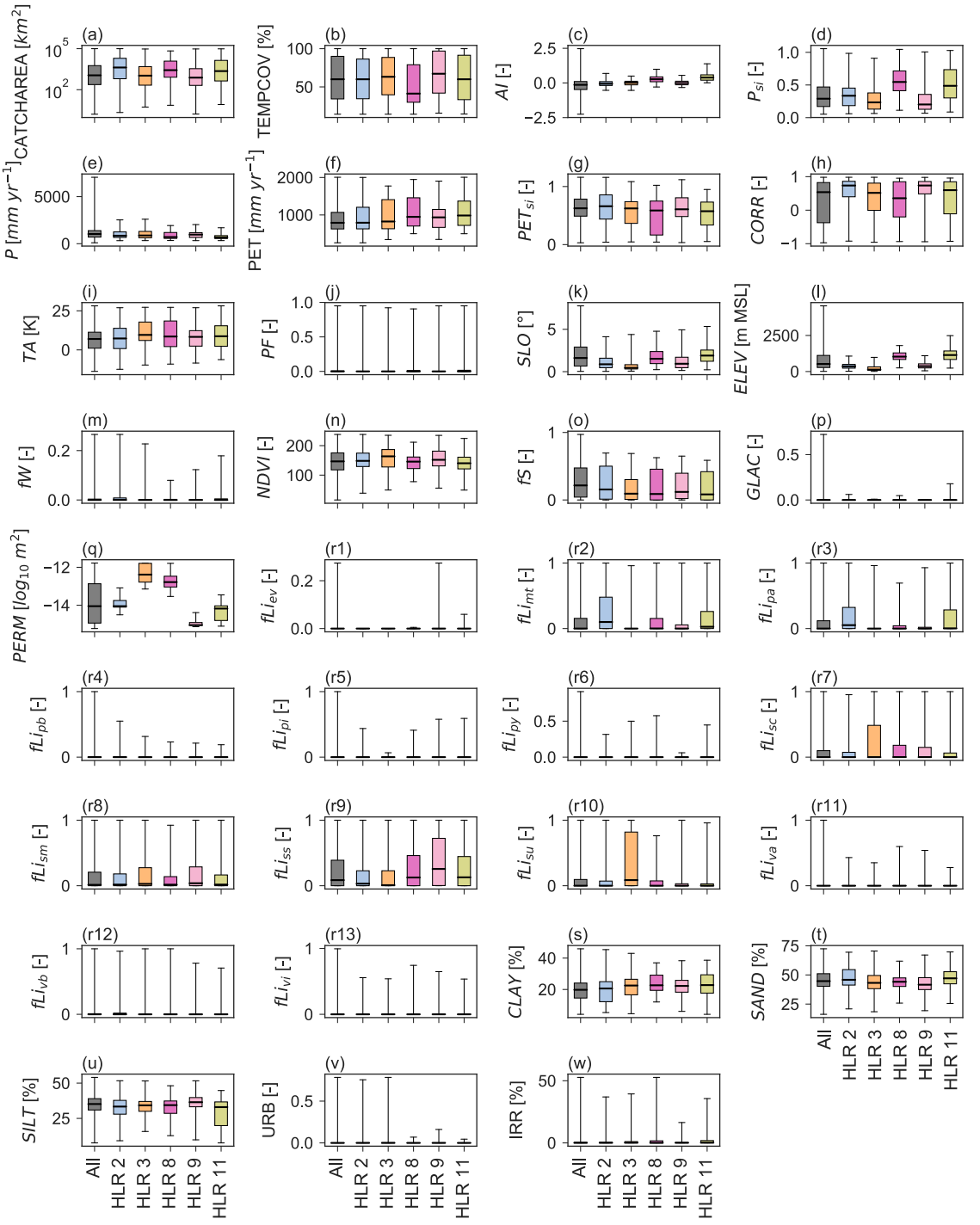


Figure A.9: Distributions of catchment area (a), temporal coverage (b), and climatic and physiographic characteristics (c-w) for the entire dataset and selected HLRs. The box plot whiskers range from the minimum to the maximum of the distribution, the box represents the inter-quartile range, and the solid line depicts the median.

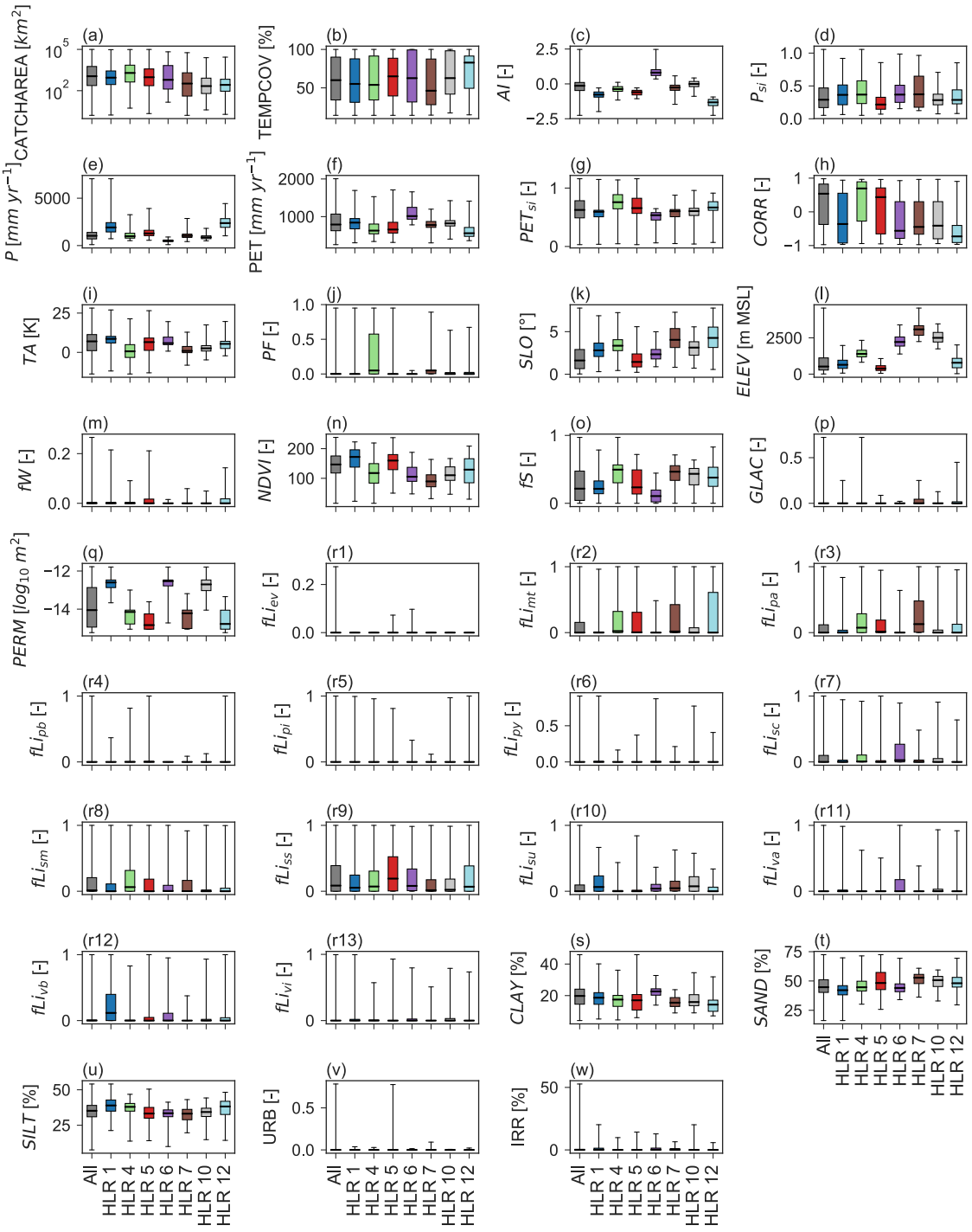


Figure A.10: Distributions of catchment area (a), temporal coverage (b), and climatic and physiographic characteristics (c-w) for the entire dataset and non-selected HLRs. Box plots as Fig. A.9.

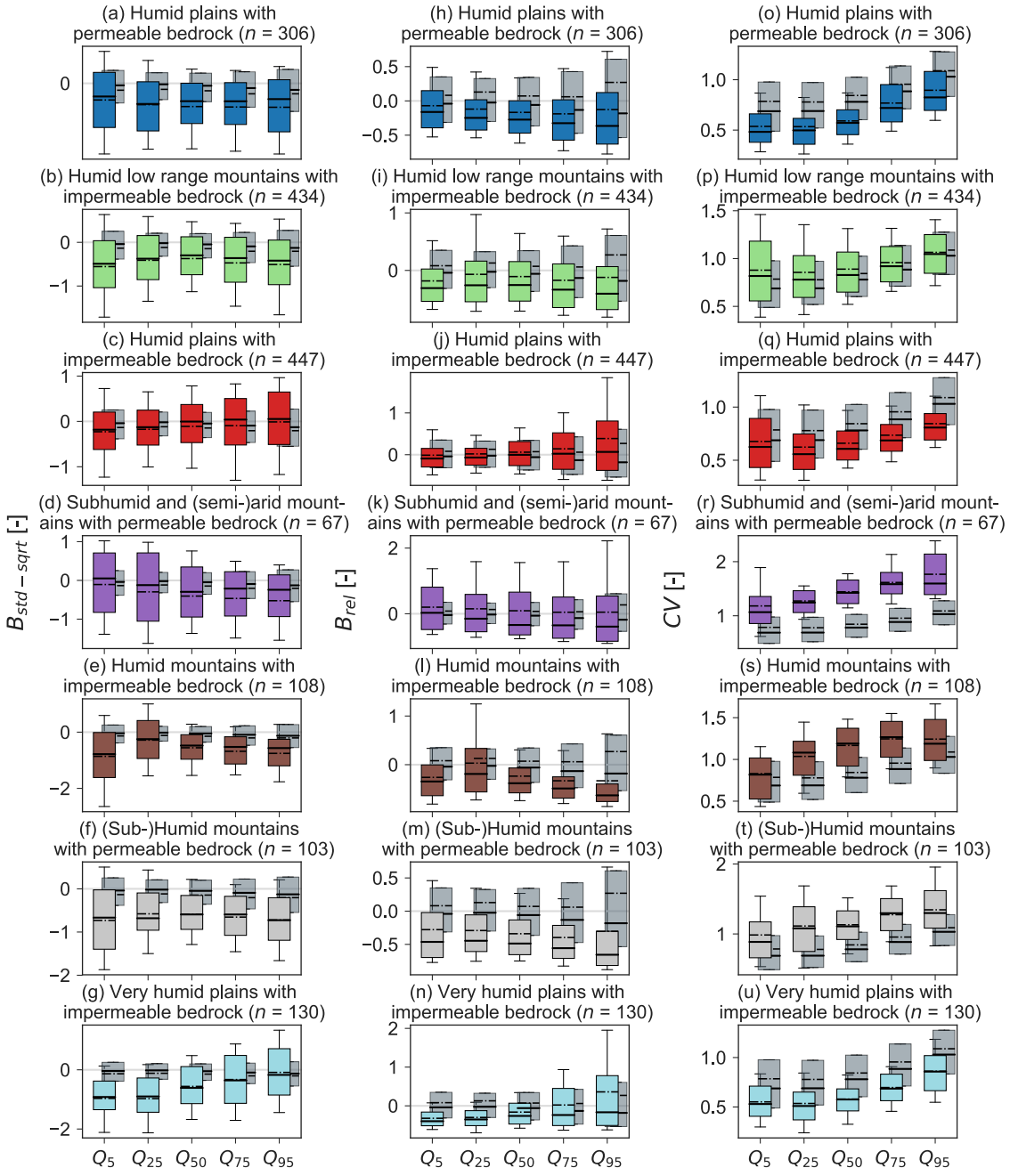


Figure A.11: Distributions of B_{rel} (a-g), $B_{std-sqrt}$ (h-n), and CV (o-u) on all five flow percentiles for the entire dataset and non-selected HLRs. n refers to the number of data points. The box plot whiskers range from the 10% to the 90% percentile of the distribution, the box represents the inter-quartile range, solid line depicts the median, and dashed line displays the mean. The dark grey boxes in the background represent the boxes of the entire dataset ($n = 3635$). For catchments locations see Fig. 4.2.

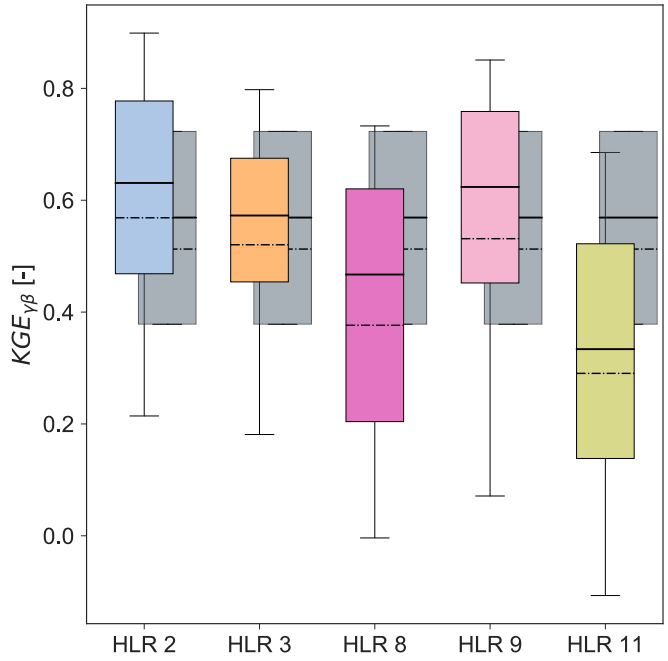


Figure A.12: Distributions of $KGE_{\gamma\beta}$ for the entire dataset and selected HLRs. The box plot whiskers range from the 10% to the 90% percentile of the distribution, the box represents the inter-quartile range, solid line depicts the median, and dashed line displays the mean. The dark grey boxes in the background represent the boxes of the the entire dataset. For catchments locations see Fig. 4.2.

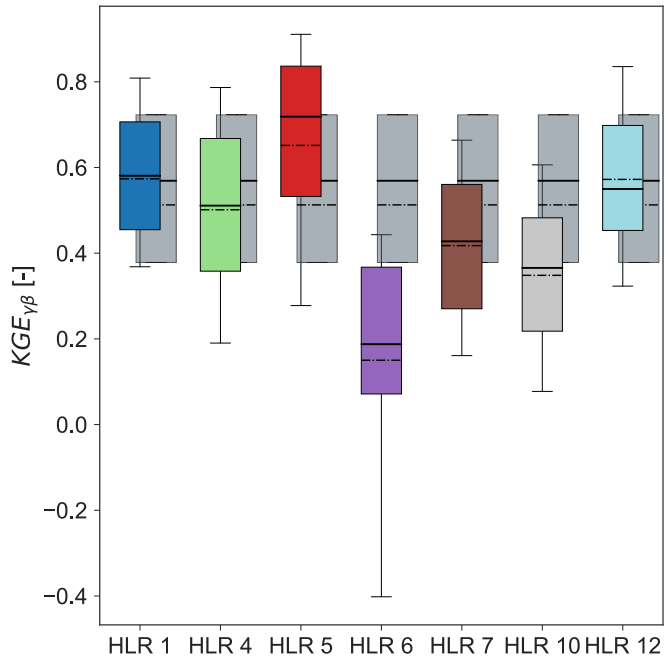


Figure A.13: Distributions of $KGE_{\gamma\beta}$ for the entire dataset and non-selected HLRs. Box plots as Fig. A.12.