The Value of Climate Forcing and Calibration for assessing Water Balance Components and Indicators of Streamflow and Total Water Storage Anomalies

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.) Physische Geographie

an der Johann Wolfgang Goethe-Universität, Frankfurt am Main

vorgelegt von
Leonie Schiebener
geb. am 06.03.1994 in Bad Soden am Taunus

Erstgutachter: Dr. H. Müller Schmied
Zweitgutachterin: Prof. Dr. P. Döll

Eingereicht 4. April 2022

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Stelle zur Prüfung eingereicht worden.

Frankfurt, den 4. April 2022

Unterschrift

Abstract

The reanalysis products and derived products, ERA5 (Copernicus Climate Change Service, 2018) and W5E5 (WATCH Forcing Data (WFD) methodology applied to ERA5) (LANGE *ET AL.*, 2021) have been recently published initiating a new phase of scientific research utilizing these datasets. ERA5 and W5E5 offer the possibility to reduce insecurities in model results through their improved quality compared to previous climate reanalyses (CUCCHI *ET AL.*, 2020). The suitability of either climate forcing as input for the hydrological model WaterGAP and the influence of the models specific calibration routine has been evaluated with four model experiments. The model was validated by analysing the models ability to produce reasonable values for global water balance components and to reproduce observed discharge in 1427 basins as well as total water storage anomalies in 143 basins using well established efficiency metrics. Bias correction of W5E5 was found to lead to more global realistic mean precipitation and consequently discharge and AET values. In an uncalibrated model setup ERA5 results in better performances across all efficiency metrics. Model results produced with W5E5 as climate input were strongly improved through calibration ultimately leading to the best performances out of all four model experiments. However, model performances considerably improved through calibration with both climate forcings hence calibration was found to have the strongest effect on model performance. Furthermore, spatial differences in performance of either forcing were identified. Snow-dominated regions show an overall better performance with ERA5, while wetter and warmer regions are better represented with W5E5. Finally, it can be concluded that W5E5 should be preferred as climate input for impact modelling; however, depending on the spatial scale and region ERA5 should at least be considered, in particular for snow-dominated regions.

Zusammenfassung

Die Veröffentlichung der Klimareanalyse, ERA5 (Copernicus Climate Change Service, 2018) und dem daraus abgeleiteten Klimadatensatz W5E5 (WATCH Force Methodik angewandt auf ERA) (LANGE *ET AL.*, 2021) eröffnen neue Forschungsmöglichkeiten. Beide Klimadatensätze bieten die Möglichkeit Unsicherheiten in den Modellergebnissen durch ihre verbesserte Qualität zu reduzieren (CUCCHI *ET AL.*, 2020). Die Eignung als Eingangsdaten für das hydrologische Modell WaterGAP sowie der Einfluss der modelleigenen Kalibrierroutine wurde mithilfe von vier Modellexperimenten überprüft. Die Validierung der Modellergebnisse erfolgte mittels der Analyse der globalen Wasserhaushaltskomponenten sowie durch die Evaluierung von Effizienzkriterien berechnet mit gemessenen Durchflusswerten in 1427 Einzugsgebieten und dem Gesamtwassergehalt in 143 Einzugsgebieten. Durch die Bias-Korrektur in W5E5 konnten belastbarere globale Niederschlagswerte und folglich auch Durchflusswerte und AET modelliert werden. Von den beiden unkalibrierten Modellexperimenten konnten die besten Ergebnisse mit ERA5 erzielt werden. W5E5 als Eingangsdatensatz für ein kalibriertes Modellexperiment führt zu den besten Modellergebnissen. Die Kalibrierung des Modells führt in beiden kalibrierten Modellexperimenten zu einer deutlichen Verbesserung der Modellergebnisse, weshalb die Kalibrierung als der größte Einflussfaktor auf gute Modellergebnisse identifiziert werden kann. Räumlich unterscheidet sich jedoch die Qualität der Modellergebnisse beider Klimadatensätze. Während ERA5 zu besseren Ergebnissen in schneedominiertem Klima führt, kann W5E5 feucht-warmes Klima besser wiedergeben. Die Analysen der Ergebnisse zeigen, dass W5E5 für die hydrologische Modellierung besser geeignet ist als ERA5. Dennoch sollte abhängig von der räumlichen Ebene und der modellierten Region, ERA5 als Klimaeingangsdatensatz zumindest in Erwägung gezogen werden.

Acknowledgement

I want to thank Dr. rer. nat. Hannes Müller Schmied and Prof. Dr. rer. nat. habil. Petra Döll for supervising this master thesis. Furthermore, I would like to thank Dr. rer. nat. Kurt Emde, who inspired me to specialise in Physical Geography and accompanied me all through my studies. I would also like to express my gratitude for the help and support I received from my family and friends.

Table of Contents

List of Figures

List of Tables

Table of Abbreviation

Abbreviation	Explanation
ADHI	African Database of Hydrometric Indices
AET	actual evapotranspiration
CFA	area correction factor
CFS	station correction factor
CRI	Coastal Resolution Improvement
CRU	Climate Research Unit
CSR	Center for Space Research
CSR-RL06	CSR GRACE mascon product
CV	coefficient of variation
DDM30	Drainage Direction Map (DÖLL and LEHNER, 2002)
ECMWF	European Centre for Medium-Range Weather Forecasts
ECPC	Experimental Climate Prediction Centres
GHM	Global Hydrological Models
GIA	Glacial isostatic adjustment
GPCC	Global Precipitation Climatology Centre
GRACE	Gravity Recovery And Climate Experiment
GRACE-FO	Gravity Recovery And Climate Experiment Follow On
GRDC	Global Runoff Data Base
GSIM	Global Streamflow Indices and Metadata archive
GSWP3	Global Soil Wetness Project Phase 3
GWSWUSE	Groundwater-Surface Water Use
HMS	Hannes Müller Schmied
IFS	Integrated Forecasting System
ISIMIP	Inter-Sectoral Impact Model Intercomparison Project
JPL	Jet Propulsion Laboratories
JPL-RL06M	JPL GRACE mascon product
KGE	Kling-Gupta efficiency index
LS	Leonie Schiebener

LWdown	downward longwave radiation
NSE	Nash-Sutcliff Efficiency
PET	potential evapotranspiration
R^2	coefficient of determination
SIEREM	Système d'Informations Environnementales sur les Ressources en Eau et leur Modélisation
SRB	Surface Radiation Budget
SWdown	downward shortwave radiation
TWSA	terrestrial water storage anomalies
WCA	actual water consumption
WFD	WATCH Forcing Data
WGHM	WaterGAP Global Hydrology Model
20CR	20^{th} Century Reanalysis

1 Introduction

Climate change in combination with a growing world population and economic development has affected the global water cycle changing water distribution and storages (DÖLL *ET AL.*, 2012). In many regions, increasing water demand frequently combined with limited resources has resulted in rivalries between different water-using sectors, e.g. agriculture, domestic use or power generation, countries as well as human interests and ecosystems (WADA *ET AL.*, 2010; United Nations, 2018). In order to sustainably manage water resources globally, quantitative estimates of the available freshwater resources and their storage distribution are indispensable. Global Hydrological Models (GHMs) aim to quantify freshwater resources with regard to spatial distribution and temporal development (MÜLLER SCHMIED *ET AL.*, 2021). They enable estimates of global water availability, sustainable use of resources (DÖLL *ET AL.*, 2012; WADA *ET AL.*, 2012), water scarcity (GOSLING and ARNELL, 2016) and groundwater depletion (WADA *ET AL.*, 2010; DÖLL *ET AL.*, 2014). A major advantage, shared with satellite altimetry, of hydrological modelling is the possibility to obtain information about freshwater resources and their storage distribution in data-limited regions where at the same time, the need for such data is often the greatest (SHEFFIELD *ET AL.*, 2018). GHMs are also used to explore the future development of freshwater resources taking human alterations of the water cycle, such as dams and irrigation projects, as well as climate change into account (SCHEWE *ET AL.*, 2014; REINECKE *ET AL.*, 2021). The findings of GHMs are used within the context of policy documents such as the International Panel for Climate Change Assessment Reports or UN World Water Reports (VELDKAMP *ET AL.*, 2018; HERSBACH *ET AL.*, 2020).

1.1 State of Research and Motivation

Despite the growing interest in GHMs and their continuing improvement over the past three decades in terms of detail, granularity, and speed as well as their increasing importance for impact studies in support of political decision-making, the quality of GHM output suffers from different uncertainty sources (VELDKAMP *ET AL.*, 2018). These sources include uncertain or simplified hydrological process representation, model structure and input data (MÜLLER SCHMIED *ET AL.*, 2014, 2021; BIERKENS *ET AL.*, 2015; SCANLON *ET AL.*, 2018). In order to judge whether these uncertainties impair the models' results to the point where they become

disinformative or implausible, the assessment of their performance in the historical period is a valuable tool. Specifically, when assessing climate impacts, good performance of a model in the historical period increases confidence in the model's future projections and uncertainties are less severe. Nevertheless, a good reproduction of climate and hydrological variables in the reference period is not a guarantee for good performance under changing climatic conditions (KRYSANOVA *ET AL.*, 2018).

Good performance of a model in the reference period however is strongly depended on the quality of the meteorological input data since it serves not only the purpose of calibration but also of model development and evaluation (KAUFFELDT *ET AL.*, 2013; MÜLLER SCHMIED *ET AL.*, 2014). Historical meteorological datasets of good quality are therefore a key element to further optimize overall quality of GHMs output and to reduce uncertainties (CUCCHI *ET AL.*, 2020). However, historical meteorological observations are unevenly distributed across the globe and data is often inconsistent. To overcome this obstacle, past observations are combined with models to generate time and space consistent reconstructions of past climate variables, so-called reanalysis (HERSBACH *ET AL.*, 2020; C3S, 2021).

The reanalysis products and derived products, further referred to as climate forcings, ERA5 (Copernicus Climate Change Service, 2018) and WFDE5 (WATCH Forcing Data (WFD) methodology applied to ERA5) (C3S, 2020) have been recently published initiating a new phase of scientific research utilizing these datasets (CUCCHI *ET AL.*, 2020). Compared to its predecessor ERA-Interim (DEE *ET AL.*, 2011) ERA5 benefits from innovations in model physics, core dynamics, and data assimilation of recent years. Apart from model improvements, ERA5 has a significantly enhanced horizontal and temporal resolution as well as an integrated uncertainty assessment (HERSBACH *ET AL.*, 2020). The WATCH Forcing Data (WFD) methodology (WEEDON *ET AL.*, 2011) has been applied to ERA5 to create the global WFDE5 meteorological forcing dataset. As WFDE5 is derived from ERA5, it inherits the same innovations compared to WFDEI (WEEDON *ET AL.*, 2014), which in turn is the resulting dataset from WFD methodology applied to ERA-Interim. In the context of WFDE5, it is worth mentioning that particularly the enhanced horizontal resolution of ERA5 leads to considerable improvement because WFD methodology requires forcing data with half-degree spatial resolution. In the case of ERA-Interim this was achieved by data interpolation and can now be obtained by data aggregation (CUCCHI *ET AL.*, 2020). Instead of using WFDE5 directly within this study, the dataset was merged with ERA5 over the oceans and temporally downscaled to daily values. This version is called W5E5 (LANGE *ET AL.*, 2021) and serves to support the bias adjustment of

climate input data for impact assessments of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) (LANGE *ET AL.*, 2021).

WFDE5 and W5E5 are used within the third simulation round of ISIMIP. The third simulation round aims to harmonize input data, focusing on impact model evaluation and improvement in ISIMIP3b protocol (ISIMIP, 2021a). Within this context, CUCCHI *ET AL.* (2020) have tested the suitability of WFDE5 for impact modelling by using it as input data for the global hydrological model WaterGAP. The model was driven by ERA5, WFDE5, and WFDEI. However, due to time and technical constraints, only uncalibrated model runs were performed and quality was assessed by comparing the resulting water balance components and evaluating model efficiency as well as river discharge seasonality for selected large river basins. The assessment revealed rather similar results for WFDE5 and WFDEI concluding that WATCH adjusted climate forcings specifically WFDE5 should be preferred as direct input compared to ERA5 (CUCCHI *ET AL.*, 2020).

To improve model output apart from utilizing better climate forcings, WaterGAP uses a basin-specific calibration routine, which matches simulated streamflow to long-term mean annual observed streamflow by adjusting the runoff coefficient and up to two additional correction factors accordingly. The remaining grid cells outside of these basins are calibrated by regionalizing the calibration factor (MÜLLER SCHMIED *ET AL.*, 2021). While calibration is routinely practiced for catchment models, WaterGAP uniquely employs this technique within the context of GHMs (DÖLL *ET AL.*, 2003; HUNGER and DÖLL, 2008; MÜLLER SCHMIED *ET AL.*, 2021). Calibration is in so far beneficial, that the fitting of simulated to observed monthly river discharge compensates to a certain degree for uncertainties regarding input data, model parameters, model structure as well as model scale and grid cell heterogeneity (MÜLLER SCHMIED *ET AL.*, 2014).

The significance of calibration for WaterGAP model results was proven by MÜLLER SCHMIED *ET AL.* (2014) who assessed the sensitivity regarding input data, model structure, human water use, and calibration. The authors concluded that the calibration method had the strongest impact on modelled freshwater fluxes and storage variations. These results are supported by the findings of KRYSANOVA *ET AL.* (2018) who evaluated the performance of regional hydrological models and GHMs with the intention of proving that models with poor performance specifically in the historical period should be excluded from climate impact studies based on ensemble means. Among other factors, KRYSANOVA *ET AL.* (2018) identified the calibration of regional

hydrological models to positively impact model performance in the historical period. They welcome the idea to include rigorous calibration approaches as used in WaterGAP. In a later study, KRYSANOVA *ET AL.* (2020) evaluated six GHMs, including WaterGAP, regarding their performance in the historical period applying common metrics with predefined thresholds. WaterGAPs superior performance was largely attributed to its calibration approach (KRYSANOVA *ET AL.*, 2020).

The above-mentioned study of MÜLLER SCHMIED *ET AL.* (2014) found the spatial differences of climate input data to be the second most influential factor regarding model output when analysed at grid cell level. Yet on a global scale these uncertainties even out. This highlights the importance of climate data specifically when analysing model performance on basin scale (MÜLLER SCHMIED *ET AL.*, 2014). However, they utilized a combinational dataset consisting of WFD and WFDEI as standard climate input which proved to be rather adverse since the radiation bias between the two datasets led to inconsistencies of actual evapotranspiration affecting storages consequently. KAUFFELDT *ET AL.* (2013) examined the consistency between climate forcing data and discharge data used for model calibration and found that screening of data should be performed prior to modelling in order to identify data-epistemic inconsistencies and to ensure that water-balance closure is possible (KAUFFELDT *ET AL.*, 2013).

Accepting inconsistencies between combinational datasets used as climate input was the price to be paid in order to increase overall model performance through calibration. Climate forcing and calibration are interlinked beyond so far presented information. Since calibration focuses on discharge, its success is tightly knit to the availability and quality of discharge data. The gauging stations currently used for calibration consist of varying time series, which reveal a severe drop in available data for the period after 1979. Hence, calibration needs to occur before 1979 requiring climate input data prior to the temporal coverage of ERA5 and W5E5, which is why artificially prolonged climate datasets are necessary. However, as presented above the prolongation has led to biases within the combinational climate datasets (MÜLLER SCHMIED *ET AL.*, 2014). LANGE (2019) developed a new method for bias adjustment and statistical downscaling for ISIMIP phase 3. Originally designed to bias adjust simulated data to observed values, it can also be applied for the bias adjustment of one climate forcing to another. LANGEs (2019) method features a more comprehensive trend preservation compared to previous methods due to the use of parametric quantile mapping. However, trend preservation causes some inhomogeneities to remain at the 1978/1979 transition (MENGEL *ET AL.*, 2021). Additionally, improvement of bias adjustment can be attributed to the newly introduced

adjustment of likelihood of individual events. The event likelihood adjustment confines extreme values to the physically plausible range and corrects imperfect distribution fits caused by the parametric quantile mapping (LANGE, 2019).

In any case, calibration may lead to serious overfitting of model results since it is targeted to only reflect one compartment of the water cycle, namely discharge. The possibility to calibrate WaterGAP regarding other output parameters has been explored for the Murray-Darling Basin by SCHUMACHER *ET AL.* (2018). One of their objectives was to test model performance in dry basins incorporating long-term hydrological trends, a field in which WaterGAP is currently lacking sufficient representation, using terrestrial water storage anomalies (TWSA) from Gravity Recovery And Climate Experiment (GRACE) satellite mission. While the model results improved regarding seasonality and trend of TWSA as well as simulation of individual water storage components, GRACE-based parameter calibration was found to be very challenging (SCHUMACHER *ET AL.*, 2018).

Nevertheless, considering parameters apart from discharge for calibration and validation as well as for evaluation of model performance is recommended (KRYSANOVA *ET AL.*, 2018; SCHUMACHER *ET AL.*, 2018). Hence, discharge and TWSA are considered here for evaluating the influence of climate forcing and calibration, which is the main objective of this master thesis. River discharge represents a unique hydrological variable given that it is the result of multiple vertical and lateral water flows within the catchment area upstream of the gauging station (HUNGER and DÖLL, 2008). Observations of river discharge are available for many regions and often comprise several decades. Yet a clear concentration of long discharge records in Europe and North America can be identified (GRDC, 2021). Measurement errors of discharge observations are considered to be relatively small compared to precipitation estimations (HUNGER and DÖLL, 2008). While the same discharge data is used for model calibration as well as for model evaluation, TWS changes derived from GRACE are entirely independent from model results. TWS is the integrated sum of all surface water, soil moisture, snow water and groundwater. It is a critical metric to monitor water supply for domestic, industrial, and agricultural sectors (TANGDAMRONGSUB *ET AL.*, 2015) but it is quite difficult to measure as well as to disaggregate the influence of individual water storages. Although very accurate, ground-based measurements only provide point-wise estimates (DORIGO *ET AL.*, 2011; TANGDAMRONGSUB *ET AL.*, 2015). Through GRACE satellite mission TWS changes can be measured with a global spatial and monthly temporal coverage. The possibility to monitor spatial and temporal variations of TWS in addition to the exclusive capability to capture

groundwater makes GRACE TWS data a very valuable resource for model calibration, validation and evaluation (EICKER *ET AL.*, 2014; TANGDAMRONGSUB *ET AL.*, 2015; SCHUMACHER *ET AL.*, 2018). When applying TWS variations for model evaluation the approach complies with KRYSANOVA *ET AL.* (2018) minimum recommended variables to be analysed and one of the two variables is independent of model results. The analysis of TWS changes recognizes the problem of overfitting and differences between simulated and measured trends and seasonality in TWS are accounted for.

1.2 Objective

In this master thesis, the analysis and comparison of discharge, TWSA, water balance components and climate variables are used to evaluate the influence of calibration and climate forcing regarding model results and performance. This master thesis aims at (1) evaluating the influence of the choice of climate forcing on water balance components (uncalibrated and calibrated model setup), (2) analysing differences between the optimal choice of climate forcing on different spatial scales (river, climate zone and global), (3) resolving whether calibration further increases the model results generated with the bias adjusted W5E5 climate forcing and (4) assessing whether W5E5 should be preferred over ERA5 for hydrological impact modelling. In chapter 2 Methodological Approach and Data the data which this master thesis is based on as well as the methods to analyse said data are presented. Furthermore, it includes a detailed description of the methodological approach to update WaterGAPs calibration station database. The results are presented in chapter three. Chapter 4 Interpretation and Discussion covers the interpretation of the results and they are critically discussed. Additionally, a future research outlook is given here. The conclusion can be found in chapter 5 Conclusion followed by the references and appendix.

2 Methodological Approach and Data

To examine the above-presented objectives a model experiment composed of four model runs was performed. The two climate forcing datasets, ERA5 and W5E5, were used as climate input data for an uncalibrated and a calibrated model run. However, both datasets had to be artificially extended to 1901 with the so-called climate forcing GSWP3, which was released within the Global Soil Wetness Project Phase 3. The data extension is necessary because WaterGAP requires a climate input dataset covering at least the time period from 1920 onwards in order to enable standard calibration. To prevent inconsistencies between GSWP3 and ERA5 and W5E5 respectively as described by KAUFFELDT *ET AL.* (2013) and experienced by MÜLLER SCHMIED *ET AL.* (2014), GSWP3 has been homogenized to both climate forcing datasets using the ISIMIP3BASD v2.5.0 quantile mapping method (LANGE, 2019, 2021).

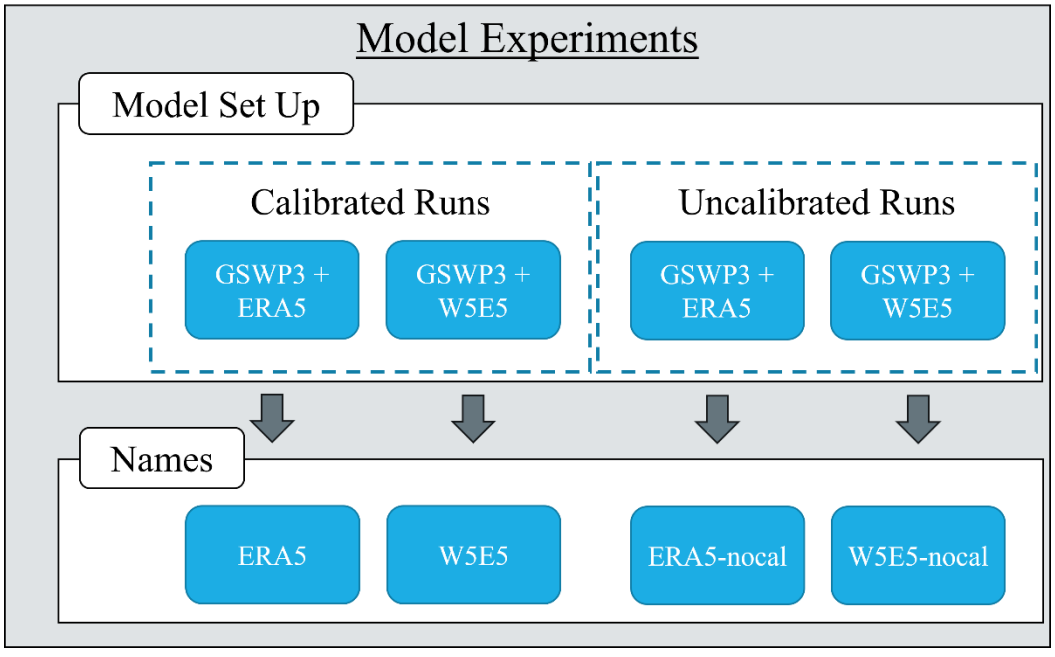

Figure 1: Schematic of experimental model set-ups with respective experiment names

Since the climate forcings of interest only date back to 1979 solely the period between 1979 and 2019 are used for analyses. Model results are compared regarding their computed values of water balance components, climate variables, discharge and TWSA. All evaluations are performed with R-Studio (see Appendix B). To evaluate model performance well-established

efficiency criteria are used which are mostly computed with the hydroGOF package (ZAMBRANO-BIGIARINI, 2020).

2.1 Climate Forcings

2.1.1 GSWP3

The GSWP3 dataset has been developed within the context of the third phase of the Global Soil Wetness Project (KIM, 2014). It offers observed data for twelve climate variables with a diurnal resolution and 0.5° x 0.5° gridded global coverage (DIRMEYER *ET AL.*, 2006; DIRMEYER, 2011). GSWP3 is based on the 20th Century Reanalysis (20CR) (COMPO *ET AL.*, 2011), which is dynamically downscaled using the spectral nudging data assimilation technique of Experimental Climate Prediction Centres (ECPCs) Global Spectral Model retaining synoptic features in the higher resolution (YOSHIMURA *ET AL.*, 2008). The spectral nudging technique is in so far advantageous that it constrains large-scale atmospheric circulation to that of observations. Additionally, 20CR data is bias-corrected with observational data from Global Precipitation Climatology Centre (GPCC) (SCHNEIDER *ET AL.*, 2014) for precipitation, Surface Radiation Budget (SRB) (STACKHOUSE *ET AL.*, 2011) for short and long wave radiation, and Climate Research Unit (CRU) TS3.23 dataset (HARRIS *ET AL.*, 2014) for monthly mean temperature and daily temperature ranges. Precipitation is further corrected by considering gauge type specific undercatch (HIRABAYASHI *ET AL.*, 2008). GSWP3 climate forcing data is available for the period between 1901 and 2014 (KIM, 2014, 2017).

The variety of data products used for bias adjustment of GSWP3 and their differing temporal availability consequently result in quality fluctuations of GSWP3 over time. However, a stabilization of data quality corresponding to quality improvements of 20CR can be identified around mid-century over the Northern Hemisphere and later over the Southern Hemisphere. Despite the improving quality of GSWP3 with advancing time, ERA5 is considered the more realistic dataset, which is why the use of GSWP3 in this study is limited to the years 1901-1979. The homogenization of GSWP3 to ERA5 and W5E5 included quantile mapping of GSWP3 time series for 1901-2004 to time series featuring the same trends but with matching distributions over the 1979-2004 reference period of the corresponding climate forcing. The trend preservation leads to residual inhomogeneities at the 1978/1979 transition, which particularly affects surface downwelling shortwave radiation over northern Europe and the Mediterranean Basin (MENGEL *ET AL.*, 2021). However, since the temporal focus of this study

is after 1979, implausible climate variables before 1979 are not expected to exceedingly impair the results.

2.1.2 ERA5

ERA5 reanalysis is the latest product of the European Centre for Medium-Range Weather Forecasts (ECMWF) covering the years from 1979 onwards. It is based on the Integrated Forecasting System (IFS) Cy41r2 with 4D-Var data assimilation (BONAVITA *ET AL.*, 2016). ERA5 has a spatial resolution of 0.25° x 0.25° and the atmospheric parameters are determined on 137 pressure levels resulting in a very fine vertical resolution. The hourly output of ERA5 is available as a preliminary version with a 5-day latency or as the final quality-checked version 2-3 months later. Additionally, ERA5 includes an uncertainty estimate obtained from the 10-member ensemble 4D-Var data assimilation system with 3-hourly output but at a coarser resolution than the original ERA5 data (HERSBACH *ET AL.*, 2020).

2.1.3 W5E5

The WATCH Forcing Data (WFD) methodology (WEEDON *ET AL.*, 2011) has been applied to the reanalysis product ERA5 to create the global WFDE5 meteorological forcing dataset. W5E5 v2.0 is a combinational dataset consisting of ERA5 over the ocean and WFDE5 v2.0 dataset over land. The eleven climate output variables are available in daily time steps. In order to align with WFDs spatial resolution, ERA5 has been aggregated to half-degree longitude-latitude grids. The data has been sequential elevation and bias corrected. CRU TS4.04 data was applied to bias correct air temperature, downward shortwave radiation and rain- and snowfall rates. Additionally, rain- and snowfall is bias-adjusted with Global Precipitation Climatology Centres GPCCv2020 monthly precipitation totals (SCHNEIDER *ET AL.*, 2011). Currently, the data is available for the period between 1979 and 2019 (LANGE *ET AL.*, 2021).

2.2 Global Hydrological Model WaterGAP

WaterGAP is a global hydrological model developed since 1996 with the purpose to quantify freshwater resources on a global scale including the impact of anthropogenic interventions. The model allows the assessment of water stress for the historic period as well as the future specifically under different climate change scenarios. Continuous model improvements have led to version two of WaterGAP with the current spatial resolution. The latest model description covers version 2.2d of WaterGAP (MÜLLER SCHMIED *ET AL.*, 2021), however for this master

thesis the most recent model version 2.2e was used. A thorough description paper is due to be published. In order to account for human water use as well as differentiating between surface and groundwater abstraction, WaterGAP comprises three substantial components: five global water use models, the linking model Groundwater-Surface Water Use (GWSWUSE), and the WaterGAP Global Hydrology Model. Consumptive water use, as the part of the abstracted water, that evapotranspirates, is computed for the sectors irrigation, livestock, domestic, manufacturing, and cooling of thermal power plants (MÜLLER SCHMIED *ET AL.*, 2021). The withdrawal water use is computed for the latter three sectors as well. GWSWUSE computes the withdrawal water use from and return flows to either surface or groundwater to obtain monthly net abstractions for both water sources (DÖLL *ET AL.*, 2012, 2014).

Finally, the WGHM simulates daily water flows and storages by computing the vertical (canopy, snow, and soil in mm) and lateral water balance (groundwater, lakes, wetlands, man-made reservoirs, and rivers in m^3). WGHM uses meteorological input data consisting of air temperature, precipitation, downward shortwave radiation, and downward longwave radiation for the computation of daily water flows. All computational steps are performed on 0.5° x 0.5° grid cells which correspond to approximately 55 x 55 km at the equator (MÜLLER SCHMIED *ET AL.*, 2021). Defined by the CRU land-sea mask (MITCHELL and JONES, 2005) the global continental area is divided into 64720 grid cells including small islands and Greenland but excluding Antarctica. Grid cells along the coastlines consist of continental and oceanic areas. The corresponding continental area is determined by subtracting the oceanic area from the total cell area. The borders between oceanic and continental area are defined by ESRIs worldmask shapefile. Continental areas in the sense of WaterGAP include land area and surface water body areas such as lakes, reservoirs and wetlands but exclude river area. The drainage direction map DDM30 (DÖLL and LEHNER, 2002) defines the upstream-downstream relation among grid cells allowing streamflow from the final water storage compartment ‘river’ to one of the eight neighbouring grid cells. Groundwater flow between gird cells does not occur (MÜLLER SCHMIED *ET AL.*, 2021) .

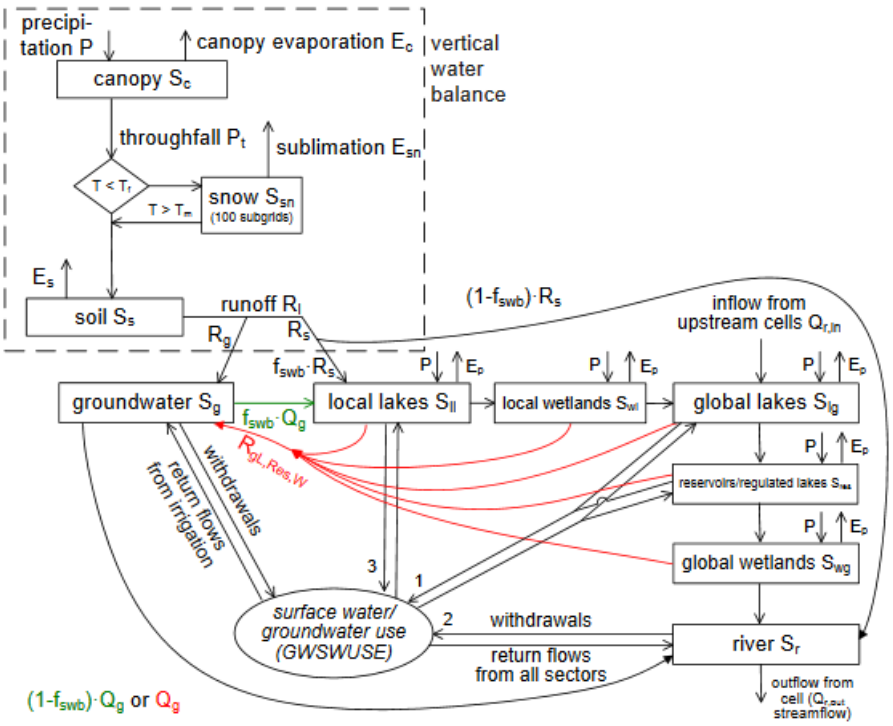

Figure 2: Schematic of WGHM in WaterGAP2.2d. Boxes represent water storage compartments. Arrows represent water flows. Green (red) colour indicate processes that occur only in grid cells with humid ((semi-)arid) climate (MÜLLER SCHMIED *ET AL.*, 2021).

2.2.1 Calibration Approach

WGHM is calibrated against observed streamflow data attempting to overcome model uncertainties regarding model parameters, input data as well as model structure including spatial resolution (DÖLL *ET AL.*, 2003; MÜLLER SCHMIED *ET AL.*, 2014). The basin-specific calibration routine matches simulated streamflow to long-term mean annual observed streamflow at now 1509 calibration stations which amounts up to 55 % of the global land area (except Antarctica and Greenland). Previous model versions used only 1319 calibration stations with approximately 54 % global coverage. The update of calibration stations was performed within the course of this master thesis. A detailed description of calibration station update and objective can be found in section 2.3 Update of Calibration Stations database. Generally, streamflow data for the calibration and evaluation of WHGM is only utilized if the respective station has an upstream area of at least 9000 km², the data covers at least four complete years, and the interstation catchment area comprises at least 30000 km². Through defining a minimum interstation area stations that are located in close proximity of each other are excluded. If available, the 30-year period from 1978 to 2009 is utilized for calibration (MÜLLER SCHMIED *ET AL.*, 2021).

Preferably, calibration is limited to the adjustment of the soil water balance through varying the runoff coefficient γ (-). The runoff coefficient together with the soil saturation $S_s/S_{s,max}$ determines the fraction of the effective precipitation (P_{eff}) that becomes runoff from land R_l, which is calculated as

$$R_l = P_{eff}\left(\frac{S_s}{S_{s,max}}\right)^{\gamma} \quad (1)$$

where S_s is the soil water storage (mm) and $S_{s.max}$ is the maximum soil water content (mm). The runoff coefficient is the only free parameter and varies between 0.1 and 5.0 in WaterGAP. The relationship between runoff as a fraction of effective precipitation and soil saturation is shown in figure 3.

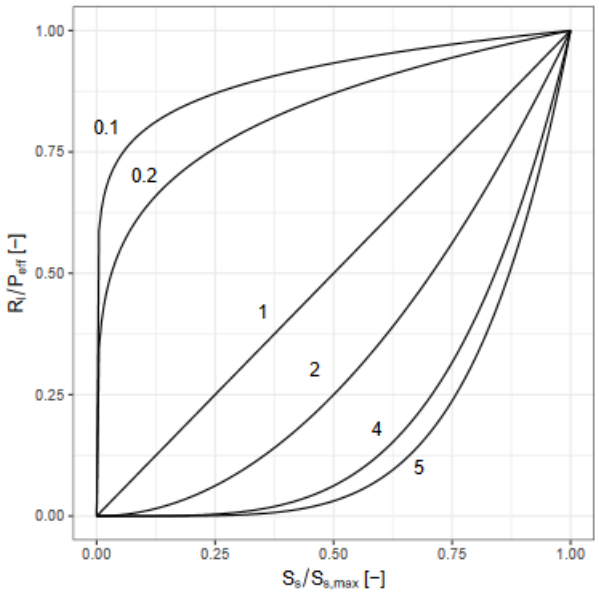

Figure 3: Relation between runoff from land and soil for different values of the runoff coefficient γ in WaterGAP (MÜLLER SCHMIED *ET AL.*, 2021)

However, in many basins adjusting the runoff coefficient alone does not lead to a satisfactory fit of simulated to observed discharge. In this case, two additional correction factors are at disposal, the area correction factor (CFA) and the station correction factor (CFS). All together, the calibration approach consists of a four-step scheme that proceeds as follows (MÜLLER SCHMIED *ET AL.*, 2021):

1. Adjustment of the runoff coefficient in the range of 0.1 and 5.0 to match simulated discharge to observed discharge within a 1 % uncertainty range
2. Adjustment of the runoff coefficient in the range of 0.1 and 5.0 to match simulated discharge to observed discharge within a 10 % uncertainty range
3. Areal correction factor (CFA): in order to conserve the mass balance, actual evapotranspiration is adjusted within the range of 0.5 and 1.5 to match simulated discharge to observed discharge within a 10 % uncertainty range
4. Station correction factor (CFS): Streamflow is multiplied in the cell where the gauging station is located by an unconstrained factor to match simulated discharge to observed discharge within a 10 % uncertainty range. Actual evapotranspiration is not adapted accordingly to avoid unphysical values, and mass is therefore not conserved.

2.2.2 Regionalization Approach

The land area outside the 1509 calibration basins benefits from the calibration due to the regionalization of the calibrated runoff coefficient. γ of so-far uncalibrated basins is adjusted within the before mentioned parameter limits and by relating the natural logarithm of γ to basin descriptors using a multiple linear regression approach. The basin descriptors consist of mean annual temperature, mean available soil water capacity, the fraction of local and global lakes and wetlands, mean basin land surface slope, the fraction of permanent snow and ice, and aquifer-related groundwater recharge factor (MÜLLER SCHMIED *ET AL.*, 2021).

2.3 Update of Calibration Stations database

During bias-adjustment of the concatenated dataset GSWP3-W5E5 for ISIMIP Phase 3a, jumps in the seasonal cycle of all climate variables were identified. The discontinuities arise at every turn of the month and are not as initially thought the result of bias-adjustment but instead inherited from GSWP3 itself (ISIMIP, 2021c). ISIMIP retraced the data in March 2021, and to this point, no update of GSWP3 has been published. In view of the long period without progress concerning GSWP3, ISIMIP decided to release the latest version of GSWP3-W5E5 for Phase

3a simulations (ISIMIP, 2021b). Consequently, the newly released GSWP3-W5E5 was used within the context of this master thesis. The original and the new dataset differs in so far that the later versions of GSWP3 (v1.09 compared to v0.5b), W5E5 (v2.0 compared to v1.0) as well as the harmonization method ISIMIP3BASD (v2.5.0 compared to v2.4.1) have been used (ISIMIP, 2021b; MENGEL *ET AL.*, 2021).

Between November and March 2021, the insecurities to whether and when a rectified version of GSWP3 would be published made the prospect to circumvent the use of the erroneous dataset ever more appealing. Apart from the problem arising specifically with GSWP3, the backwards extension of modern climate forcings, all of which begin in 1979, results in discontinuities at the 1978/1979 transition (MÜLLER SCHMIED *ET AL.*, 2016a; MENGEL *ET AL.*, 2021). This implicates that if a calibrated run should be performed, the calibration period is preferably shifted to a time frame after 1979. Since WaterGAP version 2.2d (MÜLLER SCHMIED *ET AL.*, 2021), the 30-year calibration period was, if possible, adjusted to cover the years between 1979 and 2008 instead of using the period between 1971 and 2000. However, as mentioned above, calibrating after 1979 causes a decline in available calibration stations since discharge data is scarce after 1979 for the 1319 GRDC stations used in WaterGAP 2.2d (MÜLLER SCHMIED *ET AL.*, 2021).

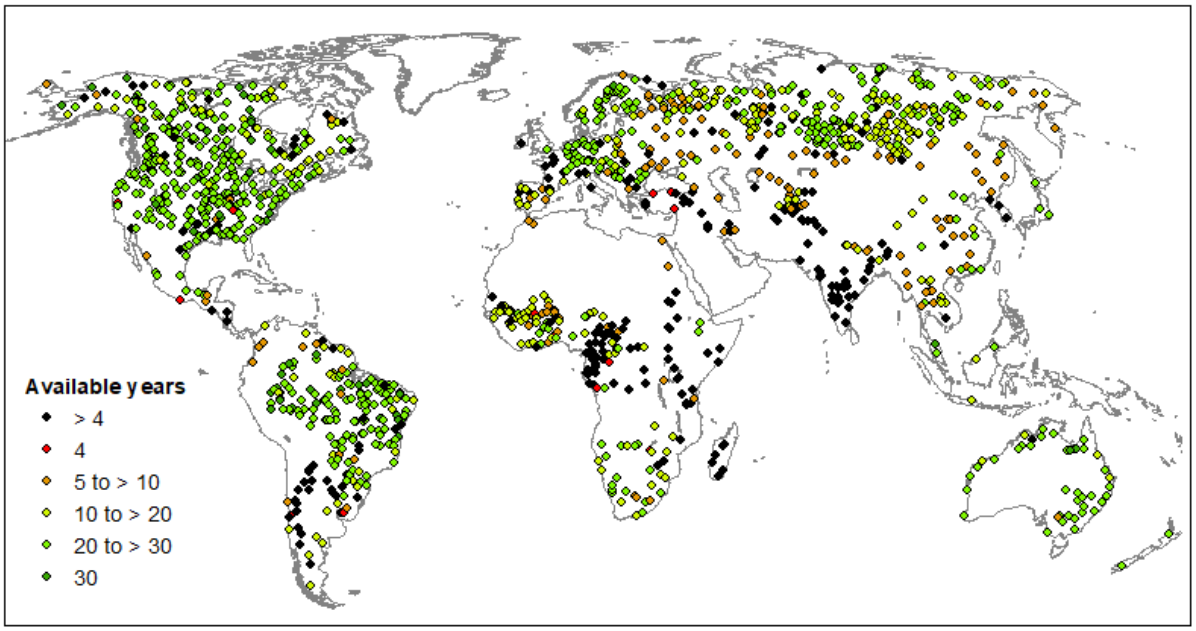

Figure 4: 1319 GRDC stations used for WaterGAP 2.2d. Stations that are lost for calibration after 1979 are indicated by black colour.

The last data retrieval from GRDC for calibration stations was conducted in 2012, which is why beginning the search for contemporary and additional discharge data at GRDC seemed

plausible. Furthermore, in recent years two other discharge databases have been published, the Global Streamflow Indices and Metadata archive (GSIM) (DO *ET AL.*, 2018; GUDMUNDSSON *ET AL.*, 2018) and the African Database of Hydrometric Indices (ADHI) (TRAMBLAY *ET AL.*, 2021). The updated and new data sources promised either the prolongation of discharge series of existing calibration stations or the enlargement of the total number of available stations meeting the calibration criteria potentially enabling successful calibration after 1979. Since including additional stations is beneficial in any way, the search for new stations was not limited to the period after 1979 but was simultaneously performed for stations containing data after 1951 and 1901. Preparing calibration for 1951 aligns with plans to calculate ERA5 from 1951 onwards, which is why it has been considered here as well.

2.3.1 Calibration Data

GRDC

The majority of streamflow data used for calibration and succeeding model evaluation is obtained from the Global Runoff Data Base (GRDB), which is maintained by the Global Runoff Data Centre (GRDC). The GRDC collects hydrological data and information on a global scale to foster scientific research in the field of climate change and risk assessment and to support water and climate-related programs of the United Nations. Daily and monthly discharge data of more than 10000 gauging stations around the globe with time series comprising up to 200 years of data are provided and perpetually updated through cooperation and exchange with national institutions, trans-national organisations and partner data centres. The data is free of charge for non-commercial users as specified by the centre's data policy and can be downloaded through the GRDC Data Portal (https://portal.grdc.bafg.de). A major effort is attributed to developing a standardised hydrological metadata profile intending to account for differences in data quality and observations used to generate the data. Universities and scientific institutions widely use hydrological data from GRDC to assess present and future freshwater resources as well as for hydrological model verification, calibration, and the validation of model results (GRDC, 2021).

GSIM

GSIM (DO *ET AL.*, 2018; GUDMUNDSSON *ET AL.*, 2018) was published in 2018 and provides a global streamflow database using publicly available data from twelve streamflow databases. More than 35000 daily discharge series have been collected and consistently formatted. Standardized processing of metadata was established to facilitate the use of discharge data for a broader community. With GSIM, the authors address issues regarding global coverage and

missing updates arising with the broadly used GRDC. Especially for South America and Asia, sufficient coverage of GRDC stations is lacking. None the less, GRDC stations and discharge series are included in GSIMs discharge catalogue. Figure 5 shows the distribution of GRDC stations and the coverage of other databases employed for the construction of GSIM. Since some databases share common spatial domains, duplicates were identified and removed during the merge of all databases leaving around 31000 discharge series. This is especially true for GRDC stations, which have been replaced by more up-to-date national databases if available. However, as described in DO *ET AL.* (2018), the replacement of GRDC with those of national databases came at the cost of losing those stations exclusive in GRDC. In addition to collecting daily discharge data, the authors of GSIM computed streamflow indices describing the respective discharge series (DO *ET AL.*, 2018; GUDMUNDSSON *ET AL.*, 2018).

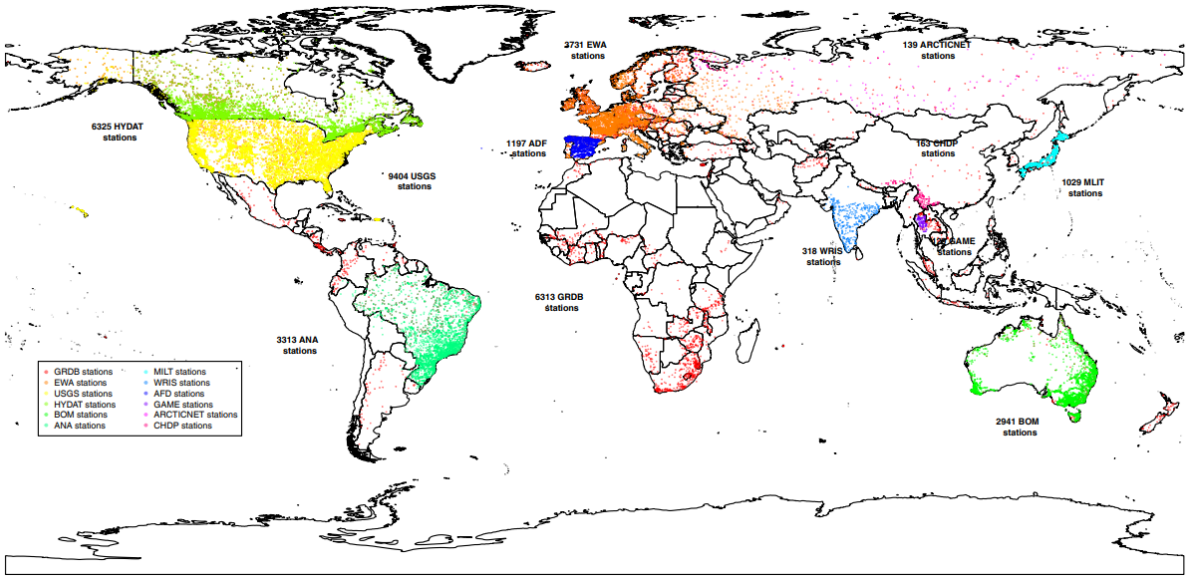

Figure 5: A schematic overview of the twelve databases used for GSIM. Further information can be found in DO *ET AL.* (2018)

ADHI

The African Database of Hydrometric Indices (TRAMBLAY *ET AL.*, 2021) was released in 2020. It provides daily discharge data, hydrological indicators, and climate variables for the African continent. The initiative's objective was to collect discharge data and metadata for the else underrepresented African continent and, in effect, complement GRDC and GSIM. The database comprises 1466 stations with daily discharge data between 1950 and 2018. For a station to be included in ADHI, it has to include a minimum of ten complete but not necessarily consecutive years in the respective period. Discharge data is retrieved from GRDC and SIEREM databases (BOYER *ET AL.*, 2006). During the merge of both databases, 106 duplicates with longer time

series in SIEREM database were identified and the GRDC duplicate was removed. In addition, a visual quality check has been performed on all discharge data. Stations showing signs of gap filling or suspicious data have been flagged in the metadata (TRAMBLAY *ET AL.*, 2021).

2.3.2 Procedure and methodological approach to update calibration database

ID Verification and Adjustment

Since the last data retrieval in 2012, GRDC has modified its station IDs, consequently stations IDs used by the working group and GRDC was mismatched. Prior to the database update, stations with new IDs needed to be identified. The search and matching process was based on two files (file with old and updated IDs, station list including metadata) provided by GRDC (GRDC, no date a, no date b). In the meantime, a new file containing updated GRDC IDs has been published. The file used for this master thesis can be found in the Appendix B.1 Calibration Station Update. Seven out of the 1319 WaterGAP calibration stations had to be identified manually because their metadata was not identical. Through visual analysis of the metadata and manual search for their updated pendant, six of these stations could be matched with an updated GRDC ID. The seventh station could not be identified, which reduced the total number of calibration stations to 1318. The respective R scripts and a list of used and produced files regarding the update of GRDC stations can be found in the Appendix B.1 Calibration Station Update.

Procedure to update the calibration station database

The procedure to update the calibration station database followed five sequential steps. All steps were performed individually for all three above-mentioned databases. Due to differences between the databases structure, spatial and temporal coverage as well as formats minor adjustments had to be made. In the first step, stations were pre-selected by exclusively analysing the metadata. Stations were selected if the corresponding catchment was larger than 9000 km² in accordance with calibration requirements and the respective discharge series ends after 1982 to ensure that the minimum available discharge series comprises four years.

In the case of GRDC, the discharge data of stations identified in step one was commissioned at the Data Centre. However, after the pre-selection process, it became evident that some newly found GRDC stations interfered with established GRDC stations in terms of interstation area. The evaluation of conflicting interstation area had to be conducted manually, given that the metadata did not allow formulating an inquiry that inspects interstation area. Therefore, GRDC

stations identified in step one were analysed in ArcMap and removed if conflicting interstation areas were identified. Whether the interstation areas of two or more stations are conflicting or not was evaluated by analysing the information regarding interstation area in the provided metadata and the station's position on DDM30. The general objective was to keep the established 1319 GRDC stations and update those if possible. In the case of conflicting interstation areas between one of the 1319 stations and a newly found station, the decision was ruled in favour of the station belonging to the 1319 stations.

During the second step the quality of the discharge series was evaluated. Monthly discharge values provided by the databases are computed from daily values. Depending on the available data points (daily values) within a month, the quality of the resulting monthly value varies. For transparency reasons the databases provided information regarding the number of days used for the respective monthly value. Based on this information a quality indicator was implemented to ensure data consistency and resilience. The threshold for missing daily values used for the computation of monthly values was set to a maximum of two days. Monthly discharge values computed from less daily values were therefore not included in any further analyses. The chosen threshold of a maximum of two missing days is a very conservative approach since the authors of ADHI recommend excluding monthly values if 5 to 10 % or more days are missing (TRAMBLAY *ET AL.*, 2021). The authors of GSIM propose to use monthly values only if they have been computed from a minimum of 25 days and denote their approach as very conservative (GUDMUNDSSON *ET AL.*, 2018).

In the third step, all discharge series of each database were merged into one file and stations were checked for complete years. If a year had less than twelve values, all values of the respective year were omitted. Based on that, stations covering at least four complete but not necessarily consecutive years after 1979 were kept. All stations with less than four years were removed and not included in any further analyses. Step three was performed for the years 1901 and 1951 as well.

The fourth step was only performed on the stations, including four complete years after 1979. The previously selected stations of all data centres were reconnected to their metadata and read into ArcMap for individual review. The visual analysis focused on identifying conflicting interstation areas between stations of the three data centres and duplicates, screening the metadata for suspicious content, and agreement of basin area in DDM30. As mentioned above, ADHI includes flagged stations containing suspicious or gap-filled data and time series with

significant regime shifts. Some of these stations have been removed during this step. In addition, some ADHI stations include the comment "donnee reconstituee" (TRAMBLAY *ET AL.*, 2021), which translates into "reconstructed data", all of which have been removed after consultation of the authors (e-mail correspondence of Hannes Müller Schmied and Yves Tramblay, 16.11.2021).

The occurrence of duplicates or multiple stations per grid cell allowed for the possibility of merging discharge series of two data sources (step five) to gain longer discharge series. Yet merging the two discharge series simply based on their matching position and metadata leaves too much ground for errors. Hence, both discharge series were displayed in one figure enabling visual analysis regarding seasonality, flow dynamics, and outliers that could stem from measurement inaccuracies or data manipulation. Apart from sharing the exact location and matching metadata, discharge series were merged if flow dynamics were similar, and seasonality seemed realistic. During the process, some discharge series showed anomalies of different forms such as continuous rise in discharge, overly high discharge values, or unnatural seasonality. For these stations, individual solutions were found. Some discharge series were modified to varying degrees. Others were removed because no solution or possible cause for the behaviour could be identified. All individual decisions and the scripts for displaying and merging station discharge data, are documented in GitHub by Hannes Müller Schmied, who developed the method and executed all tasks described in this paragraph (https://github.com/hmschmie/script-collection).

Development of 2.2e Calibration Station Database

The selection of calibration stations that are now used for calibration of WaterGAP 2.2e comprises 134 stations more than identified at the end of step five. For the construction of 2.2e calibration station database, the resulting station dataset from the above-described process was merged with stations included in 2.2d calibration. These 134 stations from 2.2d do not comprise four complete years of data after 1979, which is why they have not been considered before.

2.4 Data for Model Validation

In addition to model validation by analysing differences between simulated streamflow and the above observed discharge data, model performance is evaluated regarding its ability to compute terrestrial or total water storage (TWS). GRACE and GRACE Follow-On (GRACE-FO launched in 2018) satellite missions have measured the earth's gravity field variations since 2002 (BOERGENS *ET AL.*, 2020). TWS changes or anomalies (TWSA) can be derived from the

measured gravitational signal after subtracting the impact of atmospheric and oceanic mass variations as well as other mass variations from total gravity variations. The residual gravitational signal is attributed to changes in TWS (SCHMIDT *ET AL.*, 2006). The satellite mission consists of twin satellites following each other with a distance of approximately 220 km and flying at over 400 km altitude. When encompassing changes of the gravity field, the distance between the satellites changes in terms of length and angle. Hence, the measurement of distance changes is a proxy for gravitational changes (JET PROPULSION LABORATORY, 2021b).

TWSA are provided in so-called mass concentration blocks or "mascons" which is an alternative way of solving for gravity variations to the standard spherical harmonic approach. These TWSA mascons are directly employable monthly gridded data provided in *equivalent water thickness units* (cm). However, TWSA provided by GRACE and GRACE-FO are not relative to a measured reference value but to the 2004 to 2009 time-mean baseline. Several institutions offer mascon data for TWSA (so-called Level-3 GRACE data products) which vary in terms of data processing (Jet Propulsion Laboratory, 2021b). Two GRACE mascon solutions are used within this master thesis: the Jet Propulsion Laboratories (JPL) mascon solution RL06M.MSCNv02 (WIESE *ET AL.*, 2019) and the Center for Space Research's mascon solution CSR RL06 v02 (SAVE, 2020). At the time of writing, monthly data was available between January 2002 and April 2021. However, due to battery problems and the period between GRACE and GRACE-FO missions (GRACE: 2002-2017, GRACE-FO: May 2018- today), data incontinuities occur. Furthermore, the data features higher errors in months when the satellite's orbit is near exact-repeat, which is true for July to December 2004 and January to February 2015 (Jet Propulsion Laboratory, 2021a).

2.4.1 Mascons and Providing Institutions

RL06M.MSCNv02 – Jet Propulsion Laboratories

RL06M.MSCNv02 will further be referred to as JPL-RL06M. The dataset is based on Level-1 observations processed at JPL. C20 (degree 2 order 0) coefficients, which describe the difference between equatorial and polar radii of the equipotential surface of the Earth's gravity field, have been replaced with the solutions from Satellite Laser Ranging (CHENG *ET AL.*, 2011) due to larger uncertainties of GRACE-C20 values. Degree-1 coefficients, describing the distance between the mass centre of the Earth and its 'centre of figure', are estimated using methods from SUN, RIVA AND DITMAR (2016) and SWENSON, CHAMBERS AND WAHR (2008).

Glacial isostatic adjustment (GIA) correction has been applied according to the ICE6G-D model (RICHARD PELTIER *ET AL.*, 2018). The data is provided with a spatial resolution of 0.5° x 0.5°, however grids represent 3° x 3° equal-area cap mass concentrations which is the current native resolution of JPL-RL06M. As for WaterGAP, grids contain mixed land and ocean mass change signals. A Coastal Resolution Improvement (CRI) filter is applied to these grids separating land and ocean mass or TWSA, respectively (WIESE *ET AL.*, 2016).

CSR RL06 v02 – Center for Space Research

CSR RL06 v02 will further be referred to as CSR-RL06. The mascon has been corrected for representation on ellipsoidal earth in accordance with DITMAR (2018). C20 (degree 2 order 0) coefficients were replaced with Satellite Laser Ranging (LOOMIS *ET AL.*, 2019). C30 (degree 3 order 0) coefficients were also replaced by Satellite Laser Ranging but only for GRACE-FO. Degree-1 coefficients were corrected using estimates in Technical Note 13a (TN13a), which are derived from SUN, RIVA AND DITMAR (2016) and SWENSON, CHAMBERS AND WAHR (2008). As for JPL-RL06M, GIA correction has been applied based on ICE6G-D (RICHARD PELTIER *ET AL.*, 2018). CSR-RL06 is provided on 0.25 ° x 0.25 ° gird cells representing equal-area 1° x 1° grids, which is the current native resolution of CSR-RL06. This spatial scale has been used to comply with the newly defined hexagonal tiles, which can be split into two parts along the coast to minimize leakage between land and ocean mass signals (SAVE *ET AL.*, 2016).

2.4.2 Mascon Processing and Alignment of Mascons and WaterGAP output

The data from both mascon solutions have been manipulated to fit the World Land Mask used for WaterGAP. The re-ordered data has been aggregated over the area of 143 basins. For the JPL mascon solution the Kalman filter was used during the aggregation of liquid water equivalence. Finally, a TWSA time series for each basin was produced. In order to ensure compatibility between GRACE mascon solutions and WaterGAP output, TWSA were computed from WaterGAP TWS. In accordance with GRACE methodology, the 2004-2009 TWS baseline was calculated for each basin and consequently subtracted from every monthly TWS value.

2.5 Evaluation metrics

In hydrology, evaluation of model performance is necessary to quantitatively describe the models' ability to reproduce observed watershed behaviour. Additionally, it provides a basis for evaluating improvements regarding parameter value, model structure, and inclusion of

additional observation information as well as spatial and temporal resolution. Finally, quantitative model evaluation allows for comparisons between previous and current model results as well as across different hydrological models. KRYSANOVA *ET AL.* (2018) argue using a standardized set of evaluation metrics in order to guarantee comparability of model results. In this manner well-established evaluation metrics, such as Nash-Sutcliff efficiency (NASH and SUTCLIFFE, 1970) and Kling-Gupta efficiency index (GUPTA *ET AL.*, 2009; KLING *ET AL.*, 2012), are used within the context of this master thesis.

However, the spatial scale on which the different evaluation metrics are computed varies. Water balance components and climate variables are evaluated on a global scale excluding Antarctica and Greenland. Evaluation of streamflow indicators, Nash-Sutcliff efficiency, and Kling-Gupta Efficiency are computed for 1427 calibration basins. To evaluate the performance of TWSA, 143 larger river basins have been chosen in accordance with WaterGAP 2.2d description paper (MÜLLER SCHMIED *ET AL.*, 2021). All applied efficiency metrics and evaluation tools are described in the following section.

2.5.1 Global Parameters

Water balance components and climate variables are computed globally, and results are compared between the four model experiments. The evaluation of water balance components includes precipitation, discharge, potential evapotranspiration (PET), actual evapotranspiration (AET), net water abstraction from surface- and groundwater, consumptive water use, TWS changes, and long-term average volume balance error. In this context, computation of TWS changes is adjusted to the overall aim to evaluate satisfaction of global water balance, which is why the calculation method differs from that used to evaluate TWSA with GRACE mascon products. Here the change of TWS is obtained from the difference between the first and the last TWS value. Since the global computation of climate variables is not affected by calibration, differences in climate variables are evaluated between the two climate forcings only. In the context of this thesis, climate variables only include downward shortwave radiation, downward longwave radiation, temperature, and precipitation.

2.5.2 Streamflow Indicators

Streamflow indicators are computed to analyse the watersheds' flow characteristics and discharge dynamics. Q99 and Q90 are computed to evaluate the performance of the model in the low flow regime and the low flow extremes. Q1 and Q10 are used to analyse the ability to

reproduce high flow regimes and flood conditions. Q50 is computed to analyse the median of a discharge series. Q25 and Q75 are computed as well but are not discussed in the results and following chapters. A graphical representation of Q25 and Q75 can be found in Appendix A.

2.5.3 Efficiency Metrics

Nash-Sutcliff Efficency (NSE)

The Nash-Sutcliff (NASH and SUTCLIFFE, 1970) efficiency is a common measure for the goodness-of-fit in hydrology. It is defined as one minus the sum of the absolute squared differences between the predicted and observed values normalized by the variance of the latter. The formula reads as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (2)$$

with *S* simulated and *O* observed values. The specified variable names for simulated and observed values apply to all other equations presented here. The value of NSE can range between one and an infinitively large negative value (-∞), where 1 represents a perfect fit between simulated and observed discharge. If NSE falls below zero, the mean of the observed discharge is a better predictor than the model results. Due to the normalization of the variance of observed discharge, NSE is relatively higher in catchments with higher dynamics and lower in those with lower dynamics. Meaning that models have to perform better in catchments with lower dynamics to retrieve comparable NSE values. Problematic with NSE is that it provides no means to differentiate whether higher values result from lower mean errors or better representations of the variance (HUNGER and DÖLL, 2008). Additionally, calculating the discharge time series differences as squared values leads to an overestimation of larger values while lower ones are neglected (LEGATES and MCCABE, 1999). This leads to overestimating model performance during peak flows and an underestimation during low flow conditions.

Kling-Gupta efficiency index (KGE)

The Kling-Gupta efficiency index (GUPTA *ET AL.*, 2009; KLING *ET AL.*, 2012) has been developed to decompose NSE and solve problems associated with NSE. The incentive of KGE development was to show how the decomposition helps to enhance the understanding of overall model performance and pinpoint the reason for sub-optimal model performance through evaluation of its three sub-components. KGE shares the same value range as NSE (1 to -∞), one being the aspired value. It is computed as follows (GUPTA *ET AL.*, 2009; KLING *ET AL.*, 2012):

$$KGE = 1 - \sqrt{(KGE_r - 1)^2 + (KGE_\beta - 1)^2 + (KGE_\gamma - 1)^2} \quad (3)$$

The sub-components of KGE are the Pearson's correlation coefficient (rKGE), the bias ratio (βKGE), and the variability ratio (γKGE). The Pearson's correlation coefficient evaluates the degree of linear relationship between observed and simulated data. It ranges from -1 to 1 and indicates a perfect positive or negative linear relationship when reaching the corresponding marginal value. No linear relationship between simulated and observed values exists if the correlation coefficient is zero. βKGE describes the ratio between the mean of simulated and observed values reaching unrestrained negative or positive values. For βKGE, one is the ideal value while values above or below indicate a discrepancy between the modelled and the simulated mean. The ratio between the coefficient of variation (CV) of simulated and observed values is described by the variability ratio (γKGE), which can become infinitely negative or positive. Ideal, however is again a value of one. Values below or above one reveal the variability of the simulated values to be lesser or greater than that of the observed time series.

$$rKGE = correlation\ coefficent \quad (4)$$

$$\beta KGE = \frac{\mu_S}{\mu_O} \quad (5)$$

$$\gamma KGE = \frac{CV_S}{CV_O} = \frac{\frac{\sigma_S}{\mu_S}}{\frac{\sigma_O}{\mu_O}} \quad (6)$$

where μ denotes the mean and σ the standard deviation of the respective discharge series (GUPTA *ET AL.*, 2009; KLING *ET AL.*, 2012).

2.5.4 Evaluation metrics for TWSA

The gridded values of GRACE products are spatially averaged over the above-mentioned 143 river basins. The performance of WaterGAP is evaluated by its ability to reproduce trends and variability detected in TWSA from the two GRACE solutions. The variability of WaterGAP TWSA is evaluated using the KGE function's variability ratio. Trends in GRACE mascon solutions and WaterGAP regarding TWSA were derived from the gradient of the time series linear regression. Additionally, the coefficient of determination is used to evaluate the model's performance in replicating the dispersion of values.

Coefficient of determination

The coefficient of determination (R^2) is the square of the Pearson's correlation coefficient. Both statistics analyse the linear relationship between simulated and observed values. R^2 specifically describes the proportion of the variance of observed values that can be reproduced by the model. It ranges between zero and one, which is again the value aimed for, while zero signifies no correlation between the model results and the observed values. R^2 is calculated as:

$$R^2 = \left\{ \frac{\sum_{i=1}^{n}(O_i - \bar{O})(S_i - \bar{S})}{[\sum_{i=1}^{n}(O_i - \bar{O})^2]^{0.5}[\sum_{i=1}^{n}(S_i - \bar{S})^2]^{0.5}} \right\} \quad (7)$$

R^2 is, however quite sensitive to high extreme values and insensitive to additive and proportional differences between simulated and observed values. The same is true for the Pearson's correlation coefficient (LEGATES and MCCABE, 1999). If a model systematically over- or underpredicts all the time, R^2 can still result in values close to 1, which is why R^2 should not be considered alone when evaluating a model. To account for the coefficient of determination's inability to judge a models performance holistically, the gradient *b* of the regression on which R^2 is based should be combined with R^2, hence providing a weighted version of R^2. This is achieved by multiplying the gradient *b* with R^2. Consequently, a value of one for the gradient b is aimed for, which would in turn result in the same R^2 and weighted R^2 value (KRAUSE *ET AL.*, 2005).

3 Results

3.1 Update of Calibration Stations

3.1.1 GRDC ID Verification and Adjustment

Out of the 1319 calibration stations from WaterGAP 2.2d, 175 have changed IDs since the last data retrieval in 2012. One station, namely “Below Fort Peck” (ID: 4120901) on the Missouri river, was identified as a station with a changed ID but a new ID could not be found. As a result of backwater effects, the station included negative values and has been retracted by GRDC (e-mail correspondence between Ulrich Looser and Hannes Müller Schmied, 20.03.2019). The station has been excluded from further analyses. “Below Fort Peck” was, however reintegrated into the current calibration dataset by combining stations from WaterGAP 2.2d and stations resulting from section 2.3 Update of Calibration Stations database.

3.1.2 Discharge dataset for calibration after 1979

Of the 10361 GRDC stations, 2234 were pre-selected during the first processing step. Through visual analyses performed only on GRDC stations, another 797 were removed because of conflicting interstation area, leaving 1437 potential stations. The ADHI database includes 1466 stations, of which only 197 comprise a catchment area greater than 9000 km² and include data until 1982 or longer. 1565 of the 35000 GSIM stations met the criteria formulated in step one. As a result of excluding months failing to meet the quality criteria in step two and removing incomplete years in step three, 1199 GRDC stations, 169 ADHI stations, and 1314 GSIM stations remained for the period after 1979. The available stations for the time periods 1901 and 1951 are displayed in table 1. The visual analysis in ArcMap was performed twice (1st analysis by Leonie Schiebener, 2nd analysis by Hannes Müller Schmied) sequentially reducing the number of stations of all three data sources). After the second visual analysis, 1118 stations were selected for GRDC, and 79 ADHI stations remained. The number of GSIM stations was considerably reduced to 186 stations.

Table 1: Resulting number of stations after each processing step

Step & author	years	GRDC	ADHI	GSIM
Step 1 – LS	-	1437	197	1314
Step 2 – LS	-	as step 1, reduced discharge values for station	as step 1, reduced discharge values for station	as step 1, reduced discharge values for station
Step 3 - LS	1901	1424	189	1367
	1951	1424	189	1366
	1979	1199	169	1314
Step 4 - HMS & LS:				
1. visual analysis	1979	1143	103	521
2. visual analysis	1979	1118	79	186
Step 5 - HMS:	1979	1116	79	180

Finally, after the screening had been completed, suspicious stations had been excluded, and matching discharge series had been merged, a total of 1375 stations with four complete years after 1979 remained (step 5). The global distribution of these stations can be seen in figure 6. The total number of available years per station is colour-coded. The African continent, Asia and Russia are dominated by reddish colours, indicating that stations have relatively short discharge series. Stations with dark green signatures dominate in central Europe, Scandinavia, and the United States. Compared to figure 4 considerably large areas are lost for calibration or show fewer available stations often with shorter discharge series length.

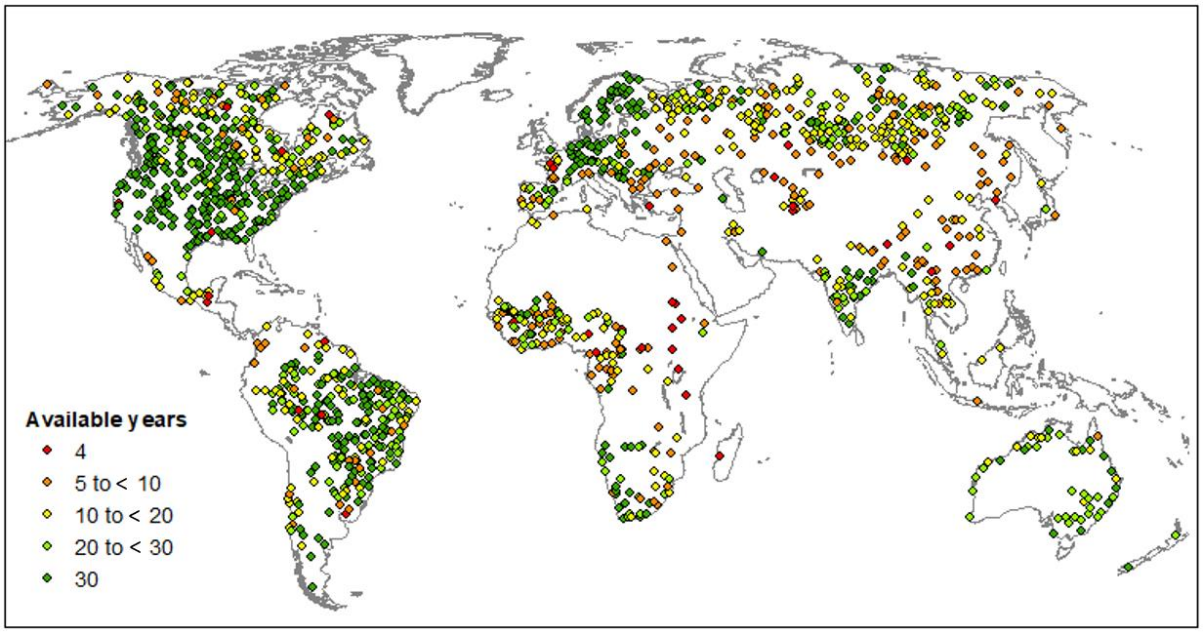

Figure 6: Resulting 1375 stations after step 5 and number of available years after 1979

3.1.3 Discharge dataset for WaterGAP 2.2e calibration

For the construction of WaterGAP 2.2e calibration dataset, the results presented above were combined with additional stations from WaterGAP 2.2d. This way, the dataset could be extended by 190 stations. In total, 1509 stations are now used for standard calibration of WaterGAP, and the calibrated area was increased by approximately 1600000 km². Figure 8 shows the affiliation to one of the three data sources. With 1252 stations, GRDC is still the dominant data source. 1109 GRDC stations are updated stations. Two of those have been merged with 2.2d stations. The resulting 143 stations were adopted from 2.2d without any alterations. In total, 177 GSIM stations are included in the calibration dataset, the majority of which are concentrated in Canada, Brazil, and India. Ten GSIM stations have been merged with 2.2d stations, and two have been merged with updated GRDC stations. ADHI contributed to the dataset with 80 stations, which are concentrated in the east and centre of the African continent. Five ADHI stations have been merged with 2.2d stations and another nine with updated GRDC stations.

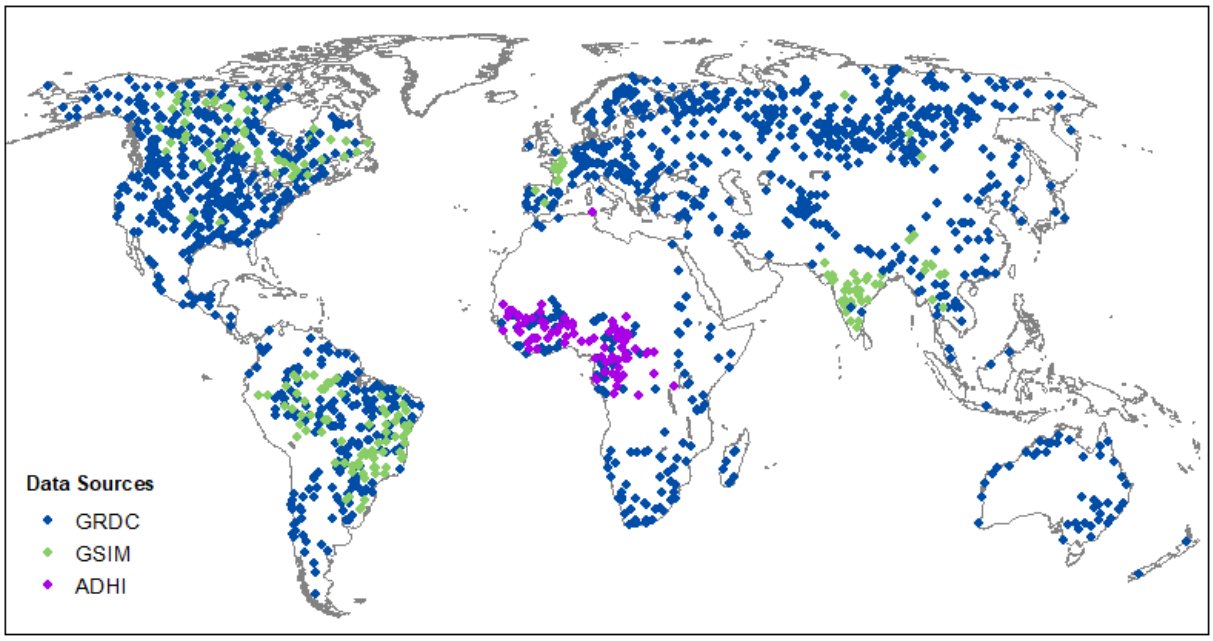

Figure 7: The final WaterGAp 2.2e calibration station dataset and the stations source database

Figure 8 shows WaterGAP 2.2e calibration stations colour-coded according to discharge data availability starting in 1912. Green is by far the most dominant colour signifying that the majority of the stations comprise 20 or more years of discharge data. Two-thirds of the 1509 stations are coloured dark green and have a discharge series equal to or longer than 30 years. A slightly higher concentration of dark green stations can be attributed to the United States, Brazil, central Europe, and Scandinavia. Stations with either 5 to 10 years or 10 to 20 years make up a neglectable quantity. Only 17 stations comprise just the minimum discharge series length of four years, most of which are located in the Middle East.

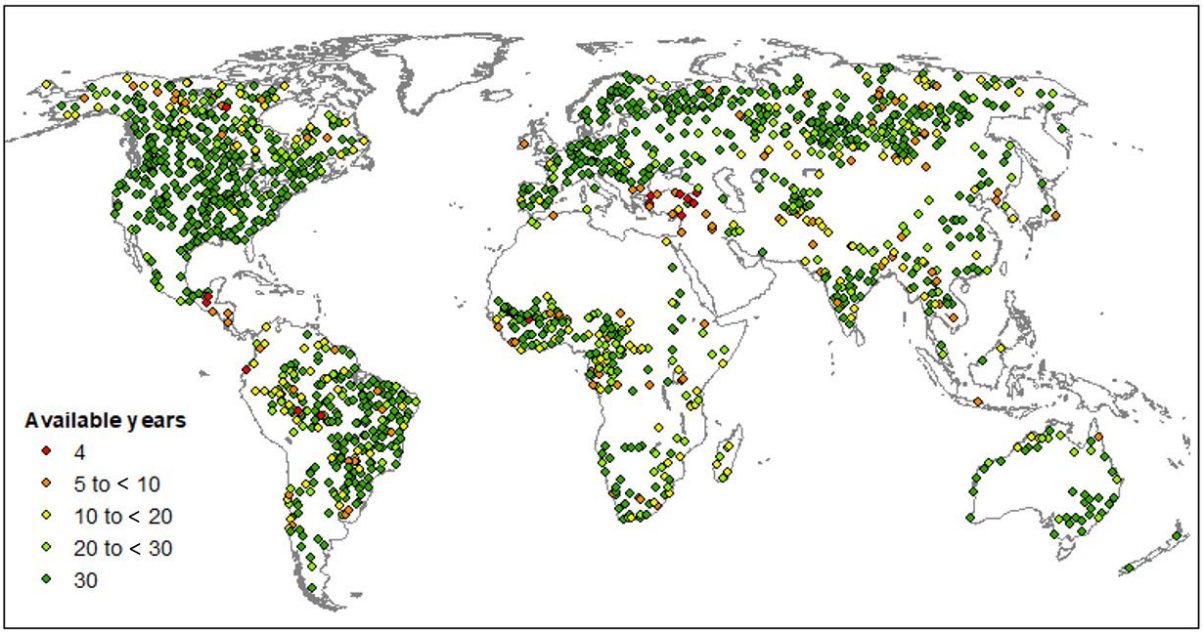

Figure 8: The final WaterGAP 2.2e calibration station dataset with availability of years

In figure 9 the number of years and stations available for calibration in relation to calibration start year is displayed. The number of available years begins to decline in the mid-1940s, but a strong exponential decrease of available years and stations alike can be identified by the end of the 1960s. The graph ends in 1979 where the number of available years has dropped by 1000 years, and the number of available stations has dropped to approximately 1300.

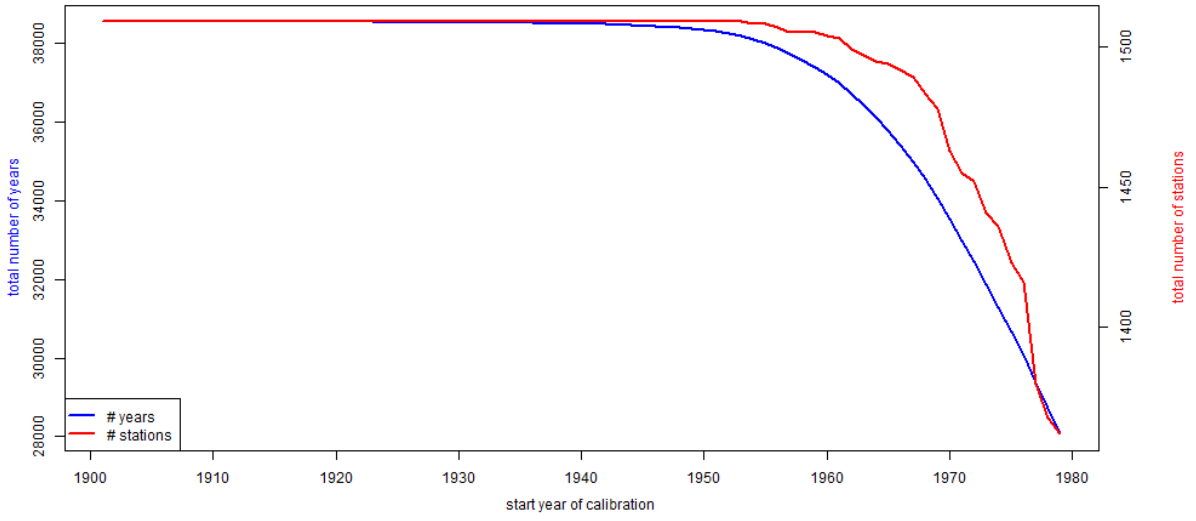

Figure 9: Number of years and stations available for calibration in relation to calibration start year

3.2 Climate Forcings and Calibration

3.2.1 Climate variables

Global climate variables have been evaluated for two climate forcings, ERA5 and W5E5. Since climate variables are unaffected by calibration analyses is limited to the two climate input datasets. Table 2 shows the mean climate variables computed over the time period of 1979 to 2019. With the exception of precipitation, all values are computed as yearly means. Precipitation is computed as the mean yearly sum. Figure 10 - 13show the corresponding spatial distribution of the climate variables. Since Greenland is not part of any other analyses, deviations between ERA5 and W5E5 in Greenland will not be discussed here.

Table 2: Climate variables for ERA5 and W5E5

Variable	ERA5	W5E5
Lwdown (W m^{-2})	323.05	323.81
Swdown (W m^{-2})	194.59	192.04
Temperature (°C)	13.58	13.65
Precipitation (mm yr^{-1})	54519489	50475474

Both forcings' yearly mean downward longwave radiation (LWdown) amounts to approximately 323 W m^{-2} with deviations on the decimal range. The overall distribution of LWdown is rather similar for both forcings (see figure 10). Significant differences in the spatial distribution of LWdown between the forcings can be identified in high latitudes (> 60° N) of Russia and North America, where weaker radiation of W5E5 stretches further to the south than for ERA5. While the weakest radiation in Russian high latitudes is largely limited to a minimum of 210 W m^{-2} in ERA5, LWdown can sink to 180 W m^{-2} over the Lena and adjacent basins in W5E5. W5E5 shows high LWdown ranging between 420 and 450 W m^{-2} over large parts of the Amazon basin. In ERA5, comparably strong LWdown is limited to small areas in the centre of the Amazon basin. Generally, W5E5 shows a higher distribution of strong LWdown (> 420 W m^{-2}) around the equator overlapping with the distribution of rainy tropical climates.

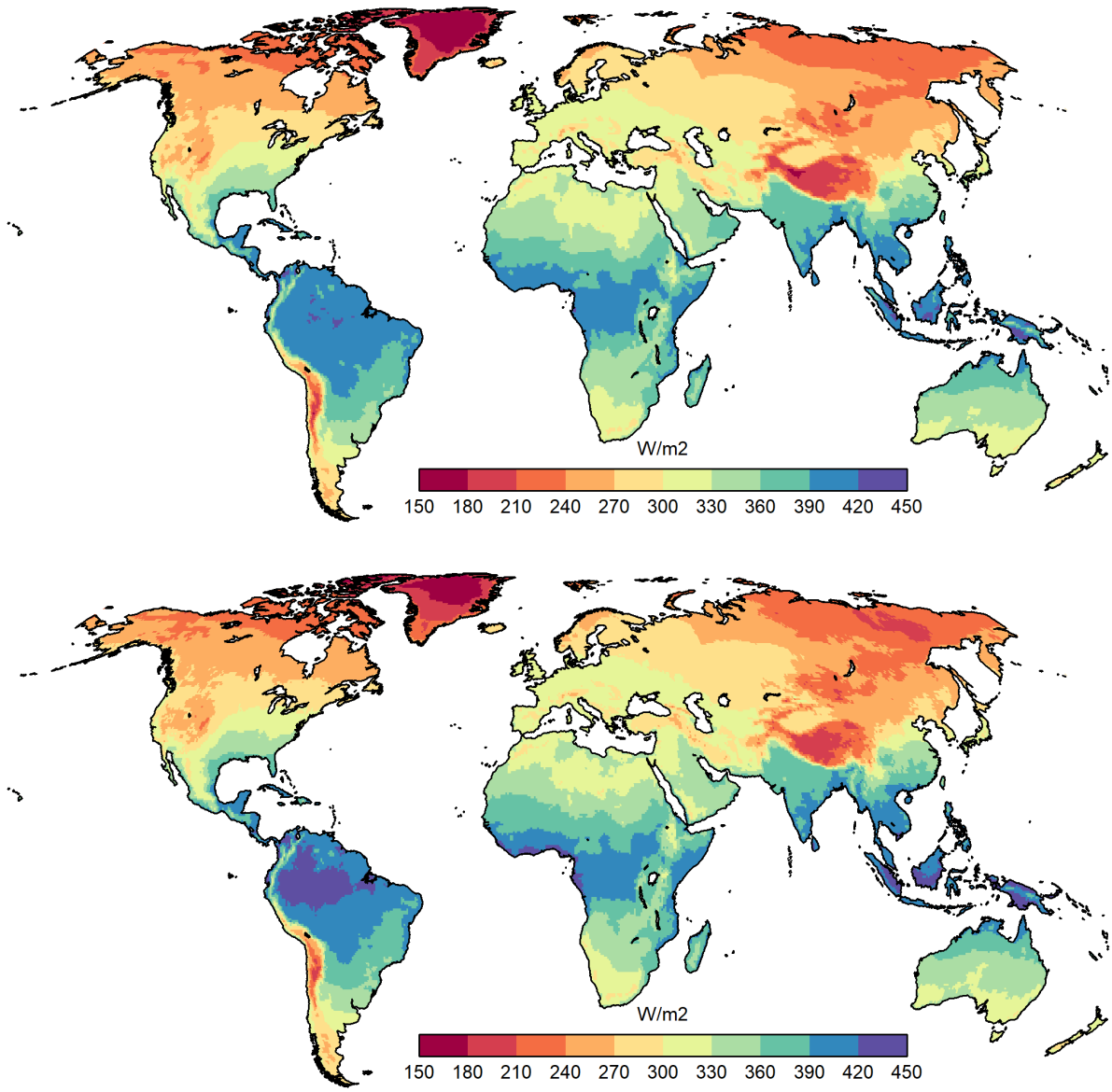

Figure 10: Mean downward longwave radiation (LWdown) between 1979 and 2019 for ERA5 (top) and W5E5 (bottom)

Mean downward shortwave radiation (SWdown) differs by approximately 2.5 W m^{-2} between the forcings. Mean SWdown amounts to 194.6 in ERA5 and 192 W m^{-2} in W5E5. Both forcings show almost horizontal distributions of SWdown, especially in the northern hemisphere (see figure 11). Weak radiation (< 60 W m^{-2}) spreads further south in W5E5 compared to ERA5. Over China, W5E5 shows a greater distribution of lower SWdown than ERA5. Differences in the distribution of high SWdown (SWdown > longtime mean) are marginal. Nevertheless, ERA5 shows a slightly higher distribution of higher SWdown values, as for example on the eastern tip of Brazil as well as the horn and centre of Africa.

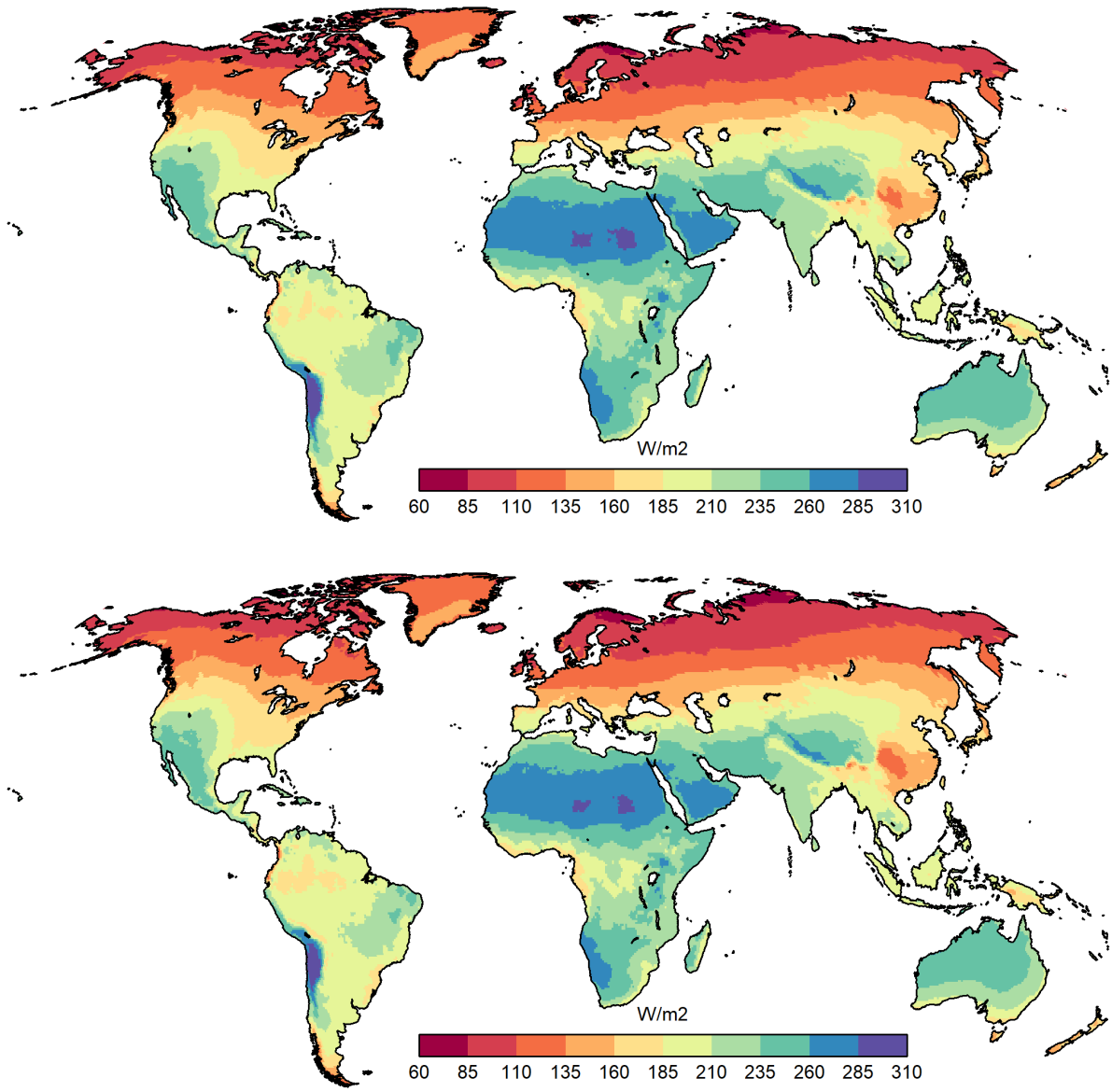

Figure 11: Mean downward shortwave radiation (SWdown) between 1979 and 2019 for ERA5 (top) and W5E5 (right)

Mean annual precipitation differs by approximately 4 Mio. mm yr^{-1}. ERA5s mean annual precipitation is considerably higher than that of W5E5. Precipitation is lower for W5E5 over Alaska and Canada but also over high latitude Russian basins, the Tibetan plateau and the Chilean Andes (climate zone E and partially D) (see figure 12). ERA5 shows lower precipitation over the eastern part of the Sahara, the Arabian peninsula as well as Australia (climate zone B). Precipitation over the Amazon basin does not seem to differ quantitatively but it differs in distributional patterns. The highest mean precipitation over the Amazon basin can be seen in the exact same location for both forcings. Yet in ERA5 higher precipitation values (2000 to 3000 mm yr^{-1}) concentrate in the centre of the basin while this precipitation class is shifted towards the east of the basin and the Brazilian coastline in W5E5. Precipitation

over central Africa is higher in W5E5 with a focus on the Congo basin. ERA5 additionally shows a greater distribution of higher precipitation values (> 1000 mm yr^{-1}) over the Himalayas and Bangladesh as well as Indonesia, Malaysia, and adjacent islands.

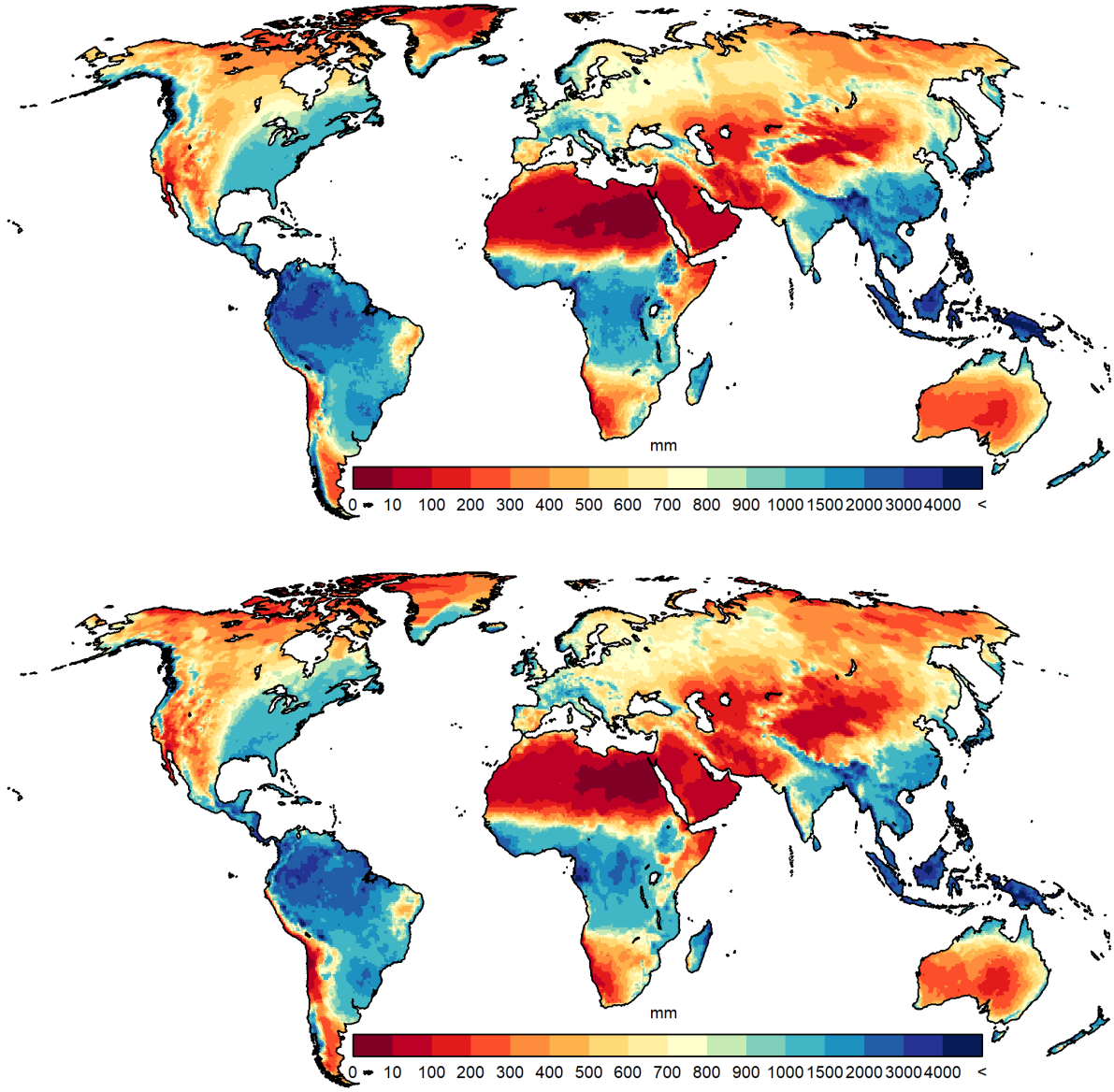

Figure 12: Mean annual precipitation between 1979 and 2019 for ERA5 (top) and W5E5 (bottom)

The mean temperature in ERA5 amounts to 13.58 °C. With a mean temperature of 13.65 °C, the climate of W5E5 is slightly warmer. Up to the -5 °C mark, lower temperatures spread further south in W5E5 than in ERA5 (see figure 13). A significantly colder area in W5E5 can be identified over the Lena basin. ERA5 shows lower mean temperatures over the Tibetan plateau. Differences in the distribution of higher temperatures between the forcings can be seen in the higher spatial coverage of temperatures above 30 °C in W5E5. Regions with greater distributions of high temperatures are the Amazon, the Sahara, India as well as Indonesia,

Malaysia and adjacent islands. On the other hand, ERA5 shows higher mean temperatures over the Arabian Peninsula.

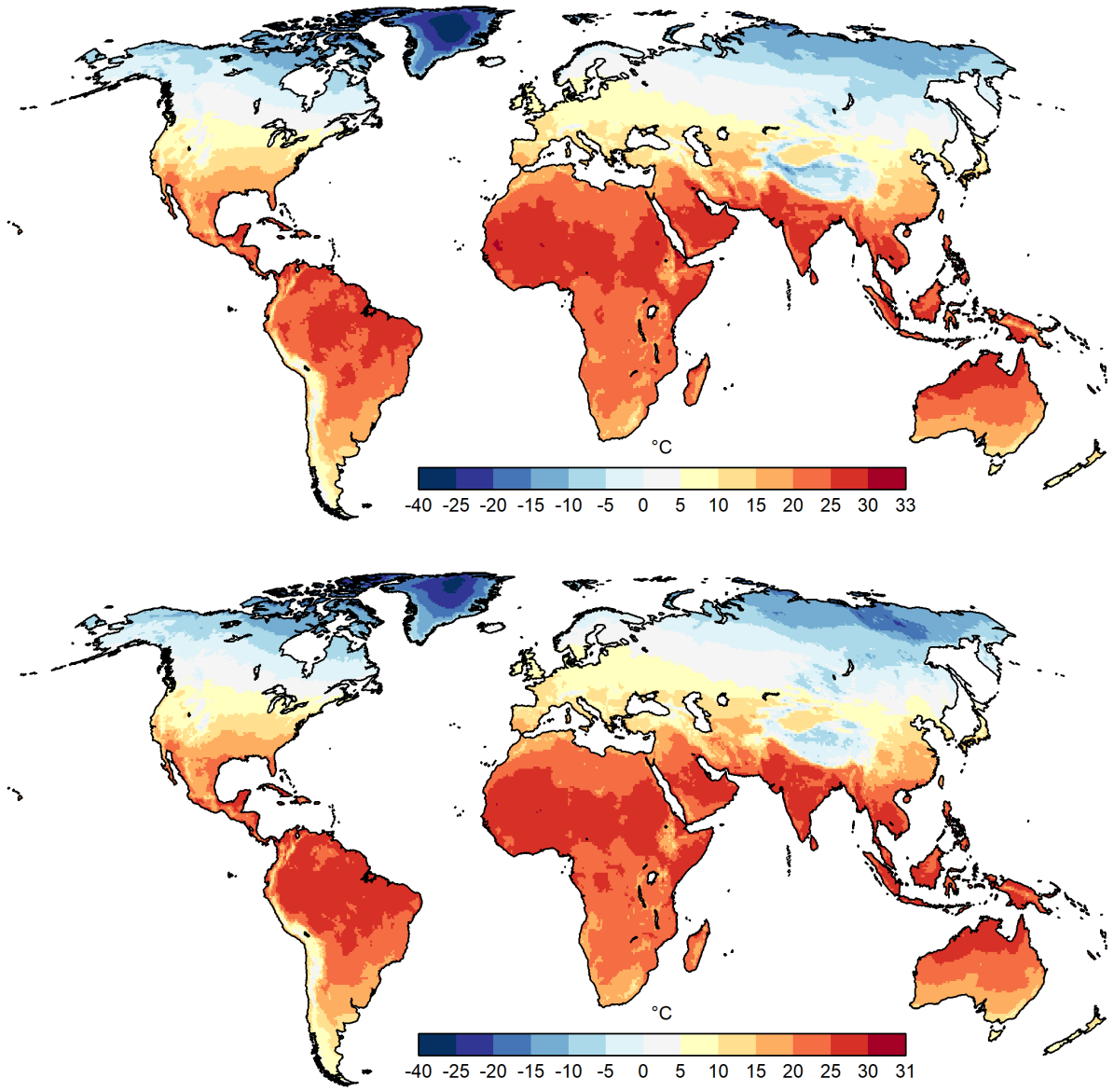

Figure 13: Mean annual temperature between 1979 and 2019 for ERA5 (top) and W5E5 (bottom)

Five major climate zones based on the Köppen-Geiger classification have been identified for both forcings. The individual climatic characteristics of the two forcings lead to differences in climate zone distribution. Figure 14 shows how the climate zones are distributed in ERA5 and W5E5.

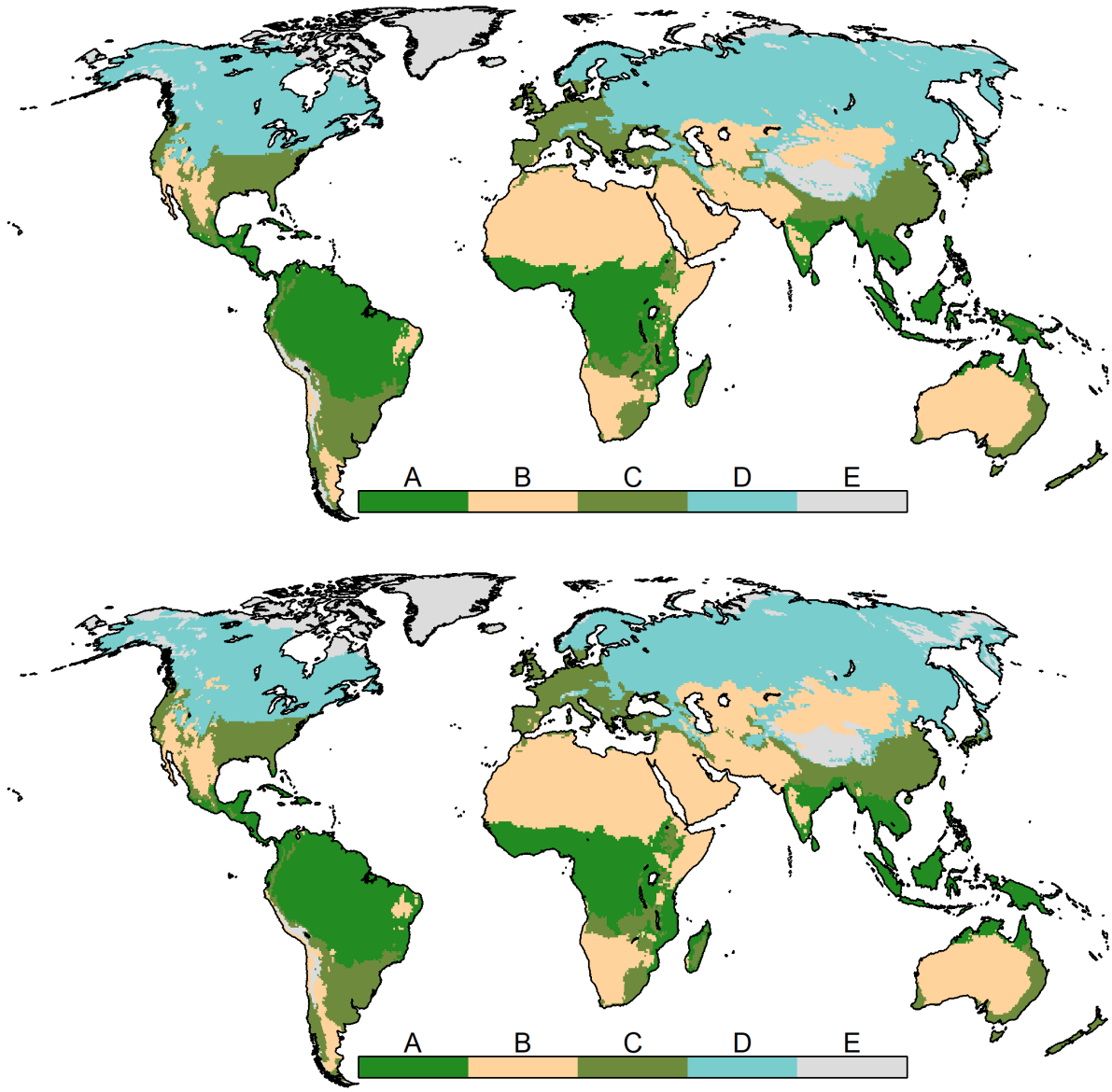

Figure 14: Climate zones according to Köppen-Geiger classification

3.2.2 Water Balance Components

Precipitation in ERA5 is approximately 8 % higher than in W5E5 (difference: 8452 km³ yr^{-1}) leading to comparably increased discharge in ERA5-nocal. Through calibration of ERA5-nocal, discharge is reduced by 4 %. Calibration of W5E5 shows a reversed trend since discharge is increased by 6 %. Discharge differences between the forcings are reduced through calibration from 4568 to 874 km³ yr^{-1}. ERA5 reveals a 1 % higher PET than W5E5 (difference: 1361 km³

yr^{-1}). The mean annual AET of ERA5-nocal is 8 % higher compared to W5E5-nocal. Through calibration, AET is increased by around 2 % in ERA5. A reversed influence of calibration on AET in W5E5 can be identified; hence AET is decreased by 3 %. Before calibration, actual consumptive use of ERA5-nocal is higher than that of W5E5-nocal (7 %). The consumptive use of ERA5 and W5E5 is decreased through calibration by 10 % and 1 %, respectively. Change of total water storage is negative for all model versions, with ERA5-nocal and W5E5-nocal showing the same value. After calibration, the negative trend of water storage of both forcings is further increased. This leads to a greater negative total water storage change in ERA5. The long-term average volume balance error is smaller than 1 $km^3\ yr^{-1}$ in all four model experiments.

Table 3: Global water balance components (excluding Antarctica and Greenland) for 1979 to 2019. All units in $km^3\ yr^{-1}$. Actual evapotranspiration includes actual consumptive water use. Actual consumptive use is the sum of row 5 and 6. Long-term average volume balance error is computed as the difference of precipitation and the sum of components 2, 4 and 8.

No.	Component	ERA5-nocal	ERA5	W5E5-nocal	W5E5
1	Precipitation	119821	119821	111370	111370
2	Streamflow into oceans and inland sinks	41892	40425	37324	39550
3	Potential evapotranspiration	150360	150359	148998	149001
4	Actual evapotranspiration	78017	79494	74133	71912
5	Actual net abstraction from surface water	1666	1497	1568	1548
6	Actual net abstraction from groundwater	-92	-79	-93	-86
7	Actual consumptive water use	1574	1418	1475	1462
8	Change of total water storage	-87	-97	-87	-93
9	Long-term average volume balance error	-0.23	-0.21	-0.21	-0.18

3.2.3 Efficiency Metrics

NSE

The NSE value aimed for is 1, values equal to or larger than 0.7 can already be judged as good model performance. Figure 15 shows the distribution of NSE values of all four model experiments. ERA5-nocal reaches NSE values larger than 0.7 in 13 % more basins than W5E5-nocal. ERA5-nocal performs superior in basins located in Siberia and former USSR territories as well as Alaska, which to a large part belong to climate zone D (see table 4). Apart from the higher latitude regions, parts of South America as well as the southeast of the United States are represented quite well by ERA5-nocal. However, the African continent, the majority of South and North America, Australia and Europe as well as the majority of China and South East Asia show NSE values below 0.1. In total, 789 basins in ERA5-nocal fall below the 0.1 NSE mark, which is about 55 % of all evaluated basins.

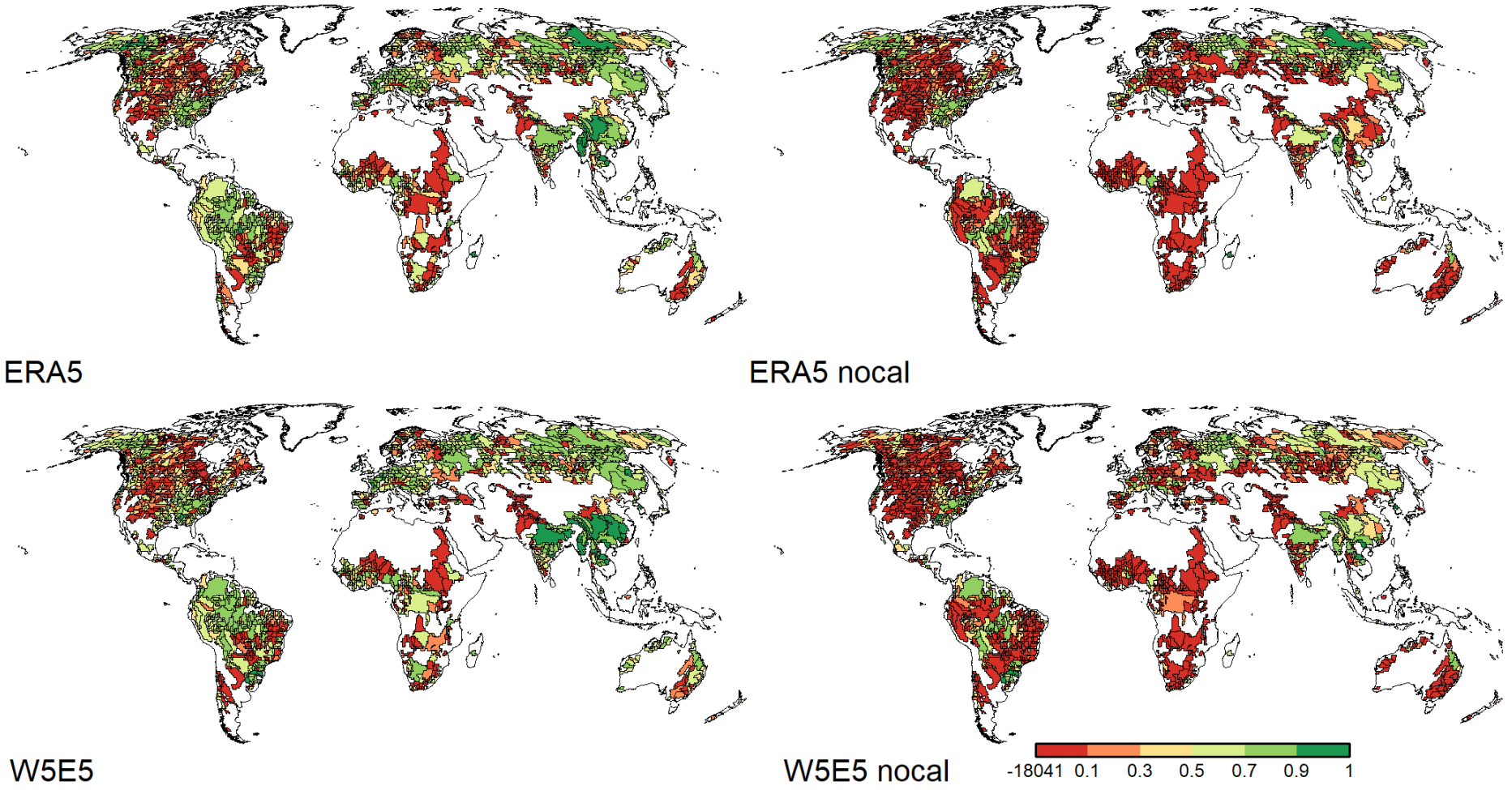

Figure 15: NSE of modelled discharge calculated for the period between 1979 and 2019 for 1427 basins

W5E5-nocal shows values between 0.7 and 0.9 for the Ganges basin and adjacent South East Asian basins as well as for western Russian basins. Values for South America are slightly better for W5E5-nocal than for ERA5-nocal. However, basins in Africa, the majority of North America, Australian as well as European basins show NSE values below 0.1. Additionally, W5E5-nocal fails to capture discharge behaviour in the majority of Russia and former USSR territories. In total, 802 basins have NSE values below 0.1, which is approximately 56 % of all evaluated basins.

Both forcings perform inferior in basins on the African continent, the majority of North America, Australia as well as Europe. More than half of the evaluated basins of both forcings have NSE values below 0.1. 77 % of ERA5-nocal and 87 % of W5E5-nocal basins reach NSE values below 0.1 in climate zone B, making it the least performing climate zone. While ERA5-nocal performs slightly better in the highest NSE class, there are only marginal differences between the two forcings in the low NSE value range.

Table 4: NSE values across climate zones for the two uncalibrated model experiments

	NSE	A	B	C	D	E	Sum
ERA5-nocal	> 0.9	2	0	2	3	1	8
	0.7 - 0.9	25	0	43	99	2	169
	0.5 - 0.7	21	5	38	108	5	177
	0.3 - 0.5	26	10	28	81	3	148
	0.1 - 0.3	29	9	34	63	1	136
	< 0.1	227	82	162	292	26	789
Sum		330	106	307	646	38	**1427**

		A	B	C	D	E	Sum
W5E5-nocal	> 0.9	1	0	8	3	0	12
	0.7 - 0.9	29	1	56	52	7	145
	0.5 - 0.7	30	7	37	88	9	171
	0.3 - 0.5	35	5	29	79	7	155
	0.1 - 0.3	26	3	24	85	4	142
	< 0.1	222	107	133	314	26	802
Sum		343	123	287	621	53	**1427**

Through calibration of ERA5, the number of basins with NSE values above 0.7 can be increased by 75 % (see table 5). Performance is significantly increased in India, South East Asia, China, Europe and the northern part of South America. Increases to values above 0.7 are mostly identified in basins located in climate zone A and C, where the number of basins has more than doubled. The Russian basins already showed relatively high values in the uncalibrated version however, these areas were further optimized by calibration. NSE values for the African continent have increased, but the continent is still dominated by values below 0.1. Comparable to the marginal performance increases on the African continent are those in North American basins. Performance of basins located in climate zone B was reduced. However, it is still the least performing climate zone with a little less than half of all basins falling below the 0.1 NSE

mark. Nevertheless, calibration reduced the number of basins with values below 0.1 by 52 %. In comparison to the 1427 evaluated basins, only one fourth is below the mark of 0.1 NSE.

Calibration of W5E5 leads to a performance increase of 146 % (number of basins with NSE values above 0.7). The most significant increase can be identified in India, South East Asia, and China, where basins show NSE values close to 1. Additionally, increases of NSE values between 0.7 and 0.9 can be identified for Russian territories (with the exception of basins surrounding the Ural mountains), the northern part of South America, Europe, and Alaskan and south-eastern U.S. basins. The strongest performance increases are located in climate Zone A where the number of basins with values above 0.7 has more than tripled. Discontinuous performance increases can be registered for the African and Australian continent, where improvements vary between 0.5 and 0.9. Only marginal improvements can be identified for central North America, which is dominated by values below 0.1. The overall number of basins with values below 0.1 could be reduced by 52 % through calibration.

Table 5: NSE values across climate zones for the two calibrated model experiments

	NSE	A	B	C	D	E	Sum
ERA5	> 0.9	3	0	10	12	1	26
	0.7 - 0.9	64	4	84	124	8	284
	0.5 - 0.7	90	15	79	153	7	344
	0.3 - 0.5	46	25	44	99	4	218
	0.1 - 0.3	35	11	30	94	8	178
	< 0.1	92	51	60	164	10	377
Sum		330	106	307	646	38	**1427**

		A	B	C	D	E	Sum
W5E5	> 0.9	2	0	21	4	4	31
	0.7 - 0.9	98	12	96	134	15	355
	0.5 - 0.7	86	25	62	145	5	323
	0.3 - 0.5	32	15	33	91	8	179
	0.1 - 0.3	32	12	23	84	7	158
	< 0.1	93	59	52	163	14	381
Sum		343	123	287	621	53	**1427**

Even though both forcings reveal significant performance increases through calibration, W5E5 performs 20% better than ERA5 (number of basins with NSE > 0.7). W5E5 is superior in India, China and South East Asia (NSE > 0.9). W5E5s superiority is less pronounced on the African and Australian continents as well as in the northern part of South America and Europe. ERA5 performs better in high latitudes of Russia. However, this observation cannot be transferred to basins outside of Russia but within the same climate zone. Slightly higher numbers of basins in climate zones D and E with values above 0.7 for W5E5 (see table 5) support the visual analyses. It is important to say that W5E5's higher performance is mostly attributed to its better performance in climate zone E. In climate zone D, W5E5 includes only two more basins with NSE above 0.7 than ERA5. Both forcings perform unsatisfactorily (NSE < 0.1) in the centre of

North America, the eastern part of South America, and the Nile and Indus basins. Additionally, both fail to represent about one-fourth of the 1427 evaluated basins (NSE values < 0.1).

KGE

When evaluating Kling-Gupta Efficiency a value of 1 is aimed for. However, values higher than 0.7 can already be judged as good performance of the model. Figure 16 shows the distribution of NSE values of all four model experiments. ERA5-nocal includes 31 % more basins with KGE values greater than 0.7 than W5E5-nocal. ERA5-nocal shows higher KGE values (0.7 < KGE < 0.9) in basins located in Siberia, Alaska and adjacent Canadian territories, which to a large part belong to climate zone D (see table 6). Additionally, ERA5-nocal performs well in the northern part of South America, including large areas of the Amazon basins and in the southeast of the United States. At the same time, ERA5-nocal shows poor performances in basins located in central North America, eastern Brazil, Australia, and on the African continent, apart from a couple of smaller basins.

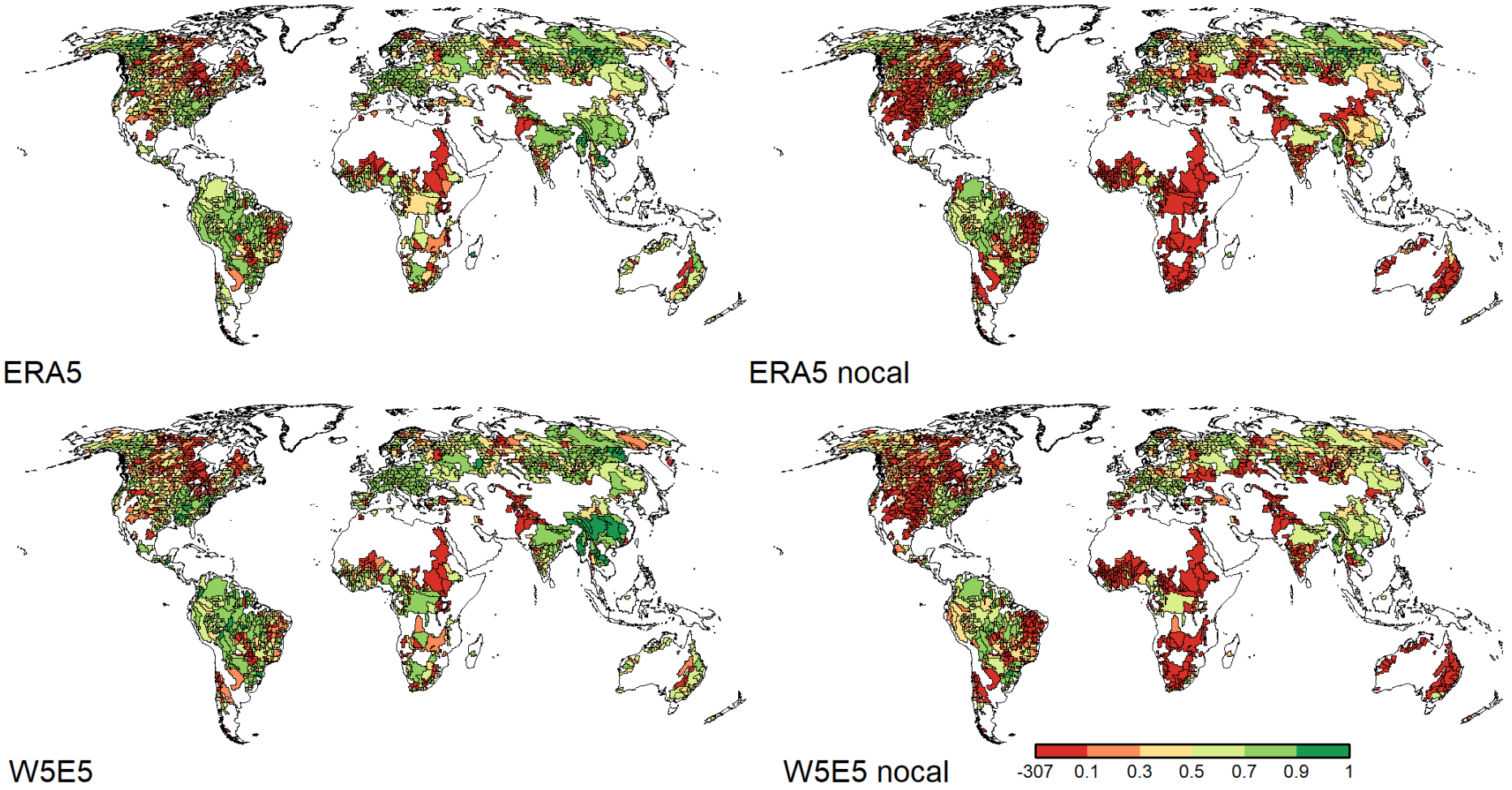

Figure 16: KGE of modelled discharge calculated for the period between 1979 and 2019 for 1427 basins

W5E5-nocal reaches a relatively good representation of basins in China and South East Asia with KGE values varying between 0.5 and 0.9. Additionally, relatively good performance is reached over the Amazon basin, the southeast of the United States as well as Europe with KGE values in the same range. However, overall performance is dominated by KGE values below 0.1 since about 41 % of all evaluated basins fall into this class. Basins with KGE values below 0.1 dominate in the centre of North America, eastern Brazil, India, Australia and on the African continent.

Even though ERA5-nocal reaches higher KGE values in more basins, they are distributed unequally across the globe. W5E5-nocal performs slightly better on the African continent than ERA5-nocal. More dominant is W5E5-nocals better representation of basins in China and South East Asia, where KGE values vary between 0.5 and 0.9. W5E5-nocal leads to basins with higher KGE values compared to ERA5-nocal in climate Zone A but ERA5-nocal includes twice as many basins with KGE values above 0.7 in climate Zone D. There are no significant differences between the performance of ERA5-nocal and W5E5-nocal in Australia, Europe and Scandinavia as well as central North America. Apart from basins in Europe, and Scandinavia, where both forcings vary between 0.1 and 0.7 without showing any particular trends, all of the formerly mentioned regions show KGE values below 0.1. Both forcings differ only marginal when analysing the number of basins with low KGE values (difference: 13 basins in favour of ERA5-nocal). The weakest performances can be identified in climate zone B, where 77 % of basins in ERA5-nocal and 89 % in W5E5 have KGE values below 0.1.

Table 6: KGE values across climate zones for the two uncalibrated model experiments

	KGE	A	B	C	D	E	Sum
ERA5-nocal	> 0.9	1	0	3	5	0	9
	0.7 - 0.9	32	1	61	101	6	201
	0.5 - 0.7	58	4	65	136	5	268
	0.3 - 0.5	59	11	46	105	5	226
	0.1 - 0.3	26	8	27	84	7	152
	< 0.1	154	82	105	215	15	571
Sum		330	106	307	646	38	**1427**

	KGE	A	B	C	D	E	Sum
W5E5-nocal	> 0.9	2	0	3	1	0	6
	0.7 - 0.9	37	1	63	48	5	154
	0.5 - 0.7	71	5	73	128	11	288
	0.3 - 0.5	40	6	40	127	10	223
	0.1 - 0.3	35	2	19	107	9	172
	< 0.1	158	109	89	210	18	584
Sum		343	123	287	621	53	**1427**

Calibration enhances the performance of ERA5 by 86 % (number of basins with KGE > 0.7) (see table 7). The number of basins with KGE values close to the optimum (KGE > 0.9) is increases by 167 %. Significant improvements can be seen in India, China, South East Asia as well as Europe and adjacent regions where values between 0.7 and 0.9 dominate. Less distinct are improvements in basins located in the centre of the United States and on the African and Australian continent. Performance in South America is further enhanced. However, that is not true for the visually prominent Orinoco River in Venezuela, where performance decreased (0.5 < KGE < 0.7). The number of basins with values below 0.1 is reduced by 58 % to 239 basins between ERA5-nocal and ERA5. Although the amount of basins with values below 0.1 in

climate zone B is reduced by more than half, it remains the climate zone where ERA5 performs the poorest.

The impact of calibration on W5E5 is even greater, leading to an increase in model performance by 187 % (number of basins with KGE > 0.7). For basins with a KGE > 0.9, an improvement of 550 % can be identified. Especially the improvements in China and South East Asia have to be highlighted. While KGE values varied between 0.5 and 0.9, with the majority ranging between 0.5 and 0.7 for W5E5-nocal, most basins show values above 0.9 for W5E5. Furthermore, W5E5 produces good values for the African continent and South America. Additionally, the high latitude regions of Russia, Alaska, and Canada are improved. The number of basins in climate zone D and E with KGE values above 0.7 has more than tripled after calibration.

Table 7: KGE values across climate zones for the two calibrated model experiments

	KGE	A	B	C	D	E	Sum
ERA5	> 0.9	4	0	7	12	1	24
	0.7 - 0.9	76	6	128	143	14	367
	0.5 - 0.7	111	23	81	181	12	408
	0.3 - 0.5	50	30	45	120	2	247
	0.1 - 0.3	30	12	18	77	5	142
	< 0.1	59	35	28	113	4	239
Sum		330	106	307	646	38	**1427**

	KGE	A	B	C	D	E	Sum
W5E5	> 0.9	10	0	19	6	4	39
	0.7 - 0.9	104	13	128	159	16	420
	0.5 - 0.7	99	28	69	155	9	360
	0.3 - 0.5	50	24	29	117	4	224
	0.1 - 0.3	22	18	14	66	7	127
	< 0.1	58	40	28	118	13	257
Sum		343	123	287	621	53	**1427**

Compared to ERA5, W5E5 has 15% more basins with KGE values above 0.7. A clear superior performance in basins located in China and South East Asia can be identified for W5E5. Higher KGE values can also be seen on the African continent. Except Russian territories where at least judging by visual analyses, ERA5 performs better, W5E5 is slightly better than ERA5. The number of basins with KGE values above 0.7 in climate zone D differs only by ten (in favour of W5E5) between both calibrated forcings. It is also worth mentioning that the number of basins with KGE values below 0.1 is lower for ERA5 than for W5E5.

Pearson's correlation coefficient (rKGE)

The optimal value for the Pearson's correlation coefficient is one. However, values equal to or larger than 0.8 already indicated good ability of the model to reproduce observed values. ERA5-nocal shows good overall performance of rKGE since basins with rKGE values above 0.8 dominate the resulting figure 17. Almost half of the evaluated basins have rKGE values above 0.8 (47 %). Basins with rKGE values below 0.5 are located in central North America and scattered across the African continent. The northern part of South America shows a coherent region with basins reaching rKGE values between 0.5 and 0.8. Regarding performance and distribution, W5E5-nocal differs only marginal from ERA5-nocal. 48 % of W5E5-nocals basins result in an rKGE value of 0.8. By visual analyses, W5E5-nocal performs superior to ERA5-nocal in Indian, Scandinavian and South American basins. A slightly better performance of W5E5-nocal can also be identified across basins on the African and Australian continents. ERA5-nocal shows a better performance in Russia and Alaska as well as the Indus basin. When partitioning the performance of both forcings across climate zones, W5E5-nocal includes more basins with good rKGE values in all climate zones with the exception of climate zone D (difference: 25 %). Both forcings fail to represent central North America, and the sub-basins of the Nile located in climate zone B. However, the number of basins with rKGE values below 0.5 is greater in W5E5-nocal.

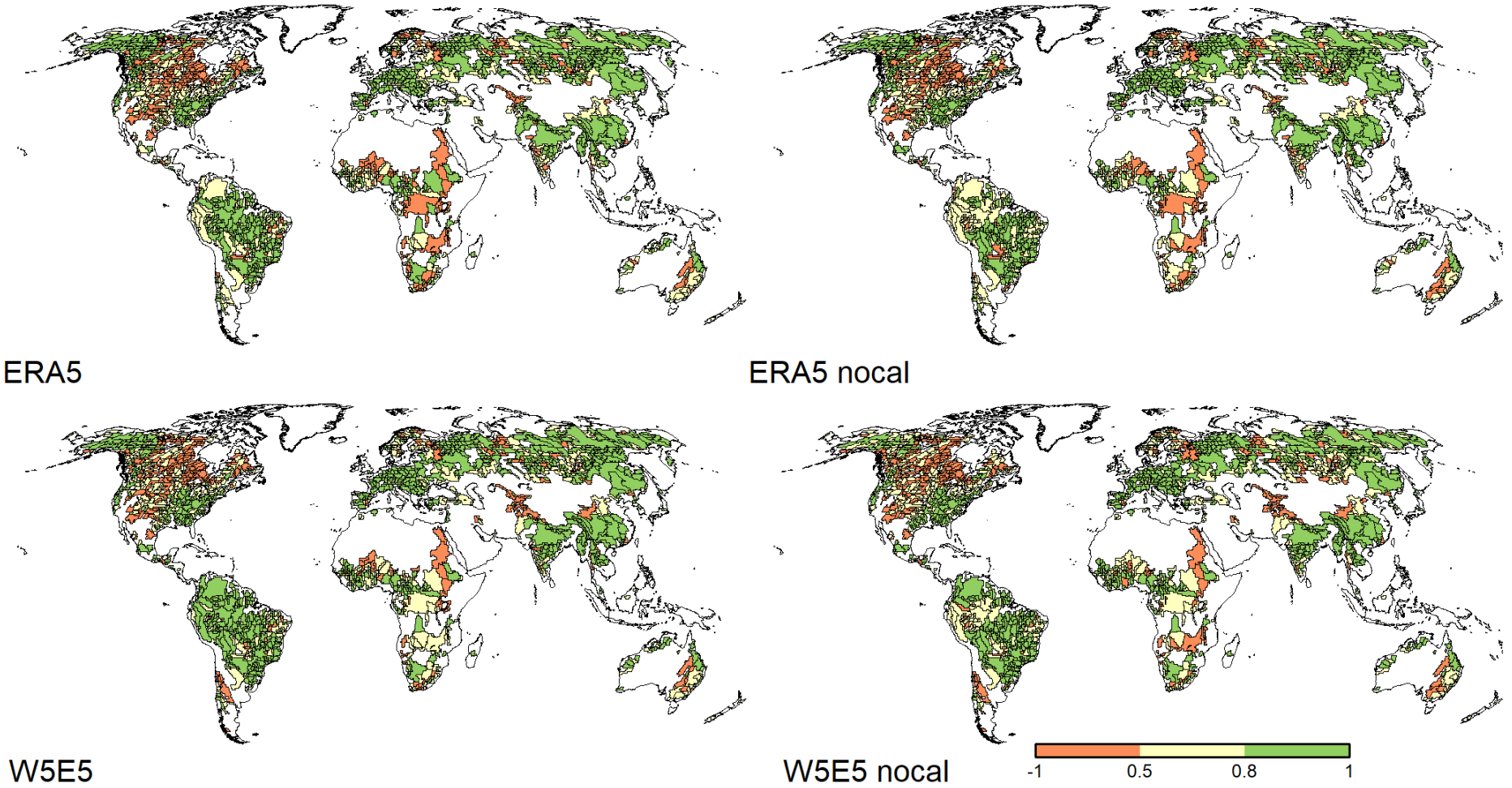

Figure 17: rKGE of modelled discharge calculated for the period between 1979 and 2019 for 1427 basins

Between ERA5-nocal and ERA5, only minimal performance increases can be identified (see figure 17). Performance increases concentrate mainly in basins on the African continent and in the northern part of South America. Some basins in Africa show deteriorating performance after calibration. All remaining basins experience close to no performance changes. However, the results presented in table 8 and 9 contradict the visual analyses since it yields a 2 % decrease in performance compared to ERA5-nocal (rKGE > 0.8). The same is true for the number of basins falling in the category with rKGE values below 0.5, which increases by 18 basins.

Table 8: rKGE values across climate zones for the two uncalibrated model experiments

	rKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
ERA5-nocal	> 0.8	163	19	185	288	20	675
	0.5 - 0.8	134	50	92	203	15	494
	< 0.5	33	37	30	155	3	258
Sum		330	106	307	646	38	**1427**

	rKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
W5E5-nocal	> 0.8	209	39	199	217	21	685
	0.5 - 0.8	107	46	69	224	14	460
	< 0.5	27	38	19	180	18	282
Sum		343	123	287	621	53	**1427**

Performance of W5E5-nocal is further increased by calibration. The number of basins with rKGE values above 0.8 increases by approximately 6 %. Figure 17 shows that performance for Russian and Alaskan basins is increased. However, the most dominant changes can be seen in South America, where with a few exceptions, all basins show values equal to 0.8 or higher. The number of basins with values below 0.5 is slightly reduced. While there are only minor performance improvements in climate zone A and E, the number of basins reaching rKGE values above 0.8 in climate zone D increases by 17 %. At the same time, the performance in climate zones B and C decreases.

Table 9: rKGE values across climate zones for the two calibrated model experiments

	rKGE	A	B	C	D	E	Sum
ERA5	> 0.8	169	16	175	279	20	659
	0.5 - 0.8	122	46	100	209	15	492
	< 0.5	39	44	32	158	3	276
Sum		330	106	307	646	38	**1427**

	rKGE	A	B	C	D	E	Sum
W5E5	> 0.8	216	35	193	253	26	723
	0.5 - 0.8	98	43	75	202	12	430
	< 0.5	29	45	19	166	15	274
Sum		343	123	287	621	53	**1427**

W5E5 contains 10 % more basins with rKGE values above 0.8 than ERA5. With the exception of climate zone D, W5E5 performs better in all climate zones. As mentioned above W5E5s performance in South American basins cannot be matched by ERA5. Even after calibration both forcings fail to represent basins in the centre of North America as well as the sub-basins of the Nile locate in climate zone B. The number of basins with values below 0.5 differs only marginally between the two forcings (differences: 2 basins).

Bias Ratio (βKGE)

The optimum value of βKGE is one, but values ranging between 0.9 and 1.1 are classified to represent a good fit between the model and observed discharge. Figure 18 shows the distribution of βKGE values of all four model experiments. 17 % of ERA5-nocal basins show βKGE in the desired class (see table 10). ERA5-nocal shows the best basin performances in the northern parts of Russia. Generally, ERA5-nocal tends to overestimate mean discharge rather than underestimating it. Underestimation of mean discharge can only be identified in basins located in Canadian territories and parts of Alaska as well as the northern parts of South America. Only 14 % of evaluated basins reach βKGE values between 0.9 and 1.1.

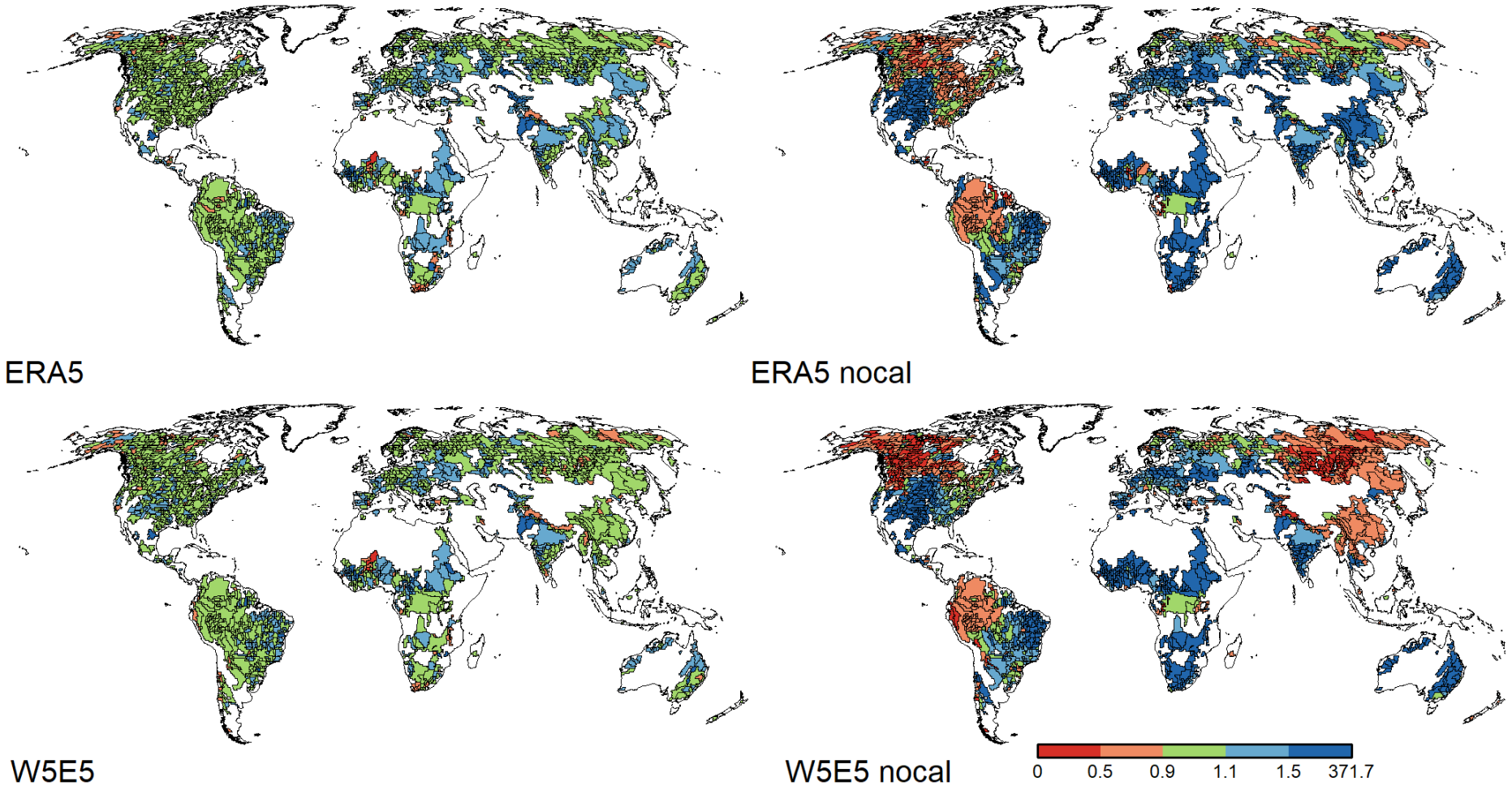

Figure 18: βKGE of modelled discharge calculated for the period between 1979 and 2019 for 1427 basins

W5E5-nocal performs the best in basins along the east coast of North America and Russian basins surrounding the Ural mountains. W5E5-nocal rather underestimates mean discharge, as large regions of North and South America as well as Russia and Asia show basins with βKGE values below 0.9. The tendency to over- or underestimate mean discharge represents the greatest difference between ERA5-nocal and W5E5-nocal. While ERA5-nocal overestimates mean discharge in Russian and Chinese basins, W5E5-nocal rather underestimates it. Across the climate zones, ERA5-nocal contains more basins with values between 0.9 and 1.1 than W5E5-nocal. The only exception is climate zone A.

Table 10: βKGE values across climate zones for the two uncalibrated model experiments

	βKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
ERA5-nocal	> 1.5	149	76	134	136	19	514
	1.1 - 1.5	65	11	90	140	12	318
	1.1 - 0.9	29	7	48	153	4	241
	0.9 - 0.5	68	8	35	187	2	300
	< 0.5	19	4		30	1	54
Sum		330	106	307	646	38	**1427**

	βKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
W5E5-nocal	> 1.5	163	105	101	83	2	454
	1.1 - 1.5	72	9	87	94	1	263
	1.1 - 0.9	36	4	44	106	3	193
	0.9 - 0.5	62	2	48	189	31	332
	< 0.5	10	3	7	149	16	185
Sum		343	123	287	621	53	1427

Calibration of ERA5 increases the number of basins falling into the desired class by 283 % (see table 10). The majority of basins in North and South America show values between 0.9 and 1.1. Significant performance increases can also be identified for basins located in Russia and

Scandinavia as well as parts of India, Australia, central Europe, and South East Asia. However, ERA5 overestimates the majority of basins outside the above-mentioned regions.

The number of basins with βKGE values between 0.9 and 1.1 in W5E5 is increased by 353% compared to W5E5-nocal. Significant performance increases can be identified for basins located in North and South America, China and South East Asia as well as central Europe, Scandinavia and Russia. Although calibration failed to increase performance in some high latitude Russian basins. On the African continent, basins with βKGE values between 0.9 and 1.1 as well as 1.1 and 1.5 are equally distributed. W5E5 tends to overestimate discharge in the remaining basins.

Table 11: βKGE values across climate zones for the two calibrated model experiments

	βKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
ERA5	> 1.5	19	6	10	8	2	45
	1.1 - 1.5	104	41	92	137	14	388
	1.1 - 0.9	188	38	197	480	20	923
	0.9 - 0.5	19	18	8	21	2	68
	< 0.5	0	3	0	0	0	3
Sum		330	106	307	646	38	**1427**

	βKGE	**A**	**B**	**C**	**D**	**E**	**Sum**
W5E5	> 1.5	15	9	6	4		34
	1.1 - 1.5	112	47	72	96	2	329
	1.1 - 0.9	203	46	193	474	32	948
	0.9 - 0.5	13	16	16	47	19	111
	< 0.5	0	5	0	0	0	5
Sum		343	123	287	621	53	**1427**

W5E5 performs approximately 3 % better than ERA5 (number of basins with 0.9 < βKGE < 1.1). W5E5 performs better in basins located in the southern Russian territories, China, South America, and on the African continent. In northern Russian basins, ERA5 performs slightly

better than W5E5. W5E5 includes more basins with optimal βKGE values in climate zones A, B, and E, whereas ERA5 includes slightly more basins in climate zones C and D.

Variability Ratio (γKGE)

If the variability ratio (γKGE) reaches a value of one, the model's optimal representation of discharge variability can be assumed. Nevertheless, γKGE values between 0.9 and 1.1 already signify good model performance. ERA5-nocal shows rather adverse performance of γKGE. Satisfactorily performing basins are scares all over the globe and distributed unequally. 20 % of the 1427 evaluated basins show γKGE values between 0.9 and 1.1. There is no significant dominance of a specific climate zone. W5E5-nocal leads to good performance in China, central Europe and the Amazon basin. Nevertheless, good performing basins are scares and only 19 % of all evaluated basins reach γKGE values between 0.9 and 1.1.

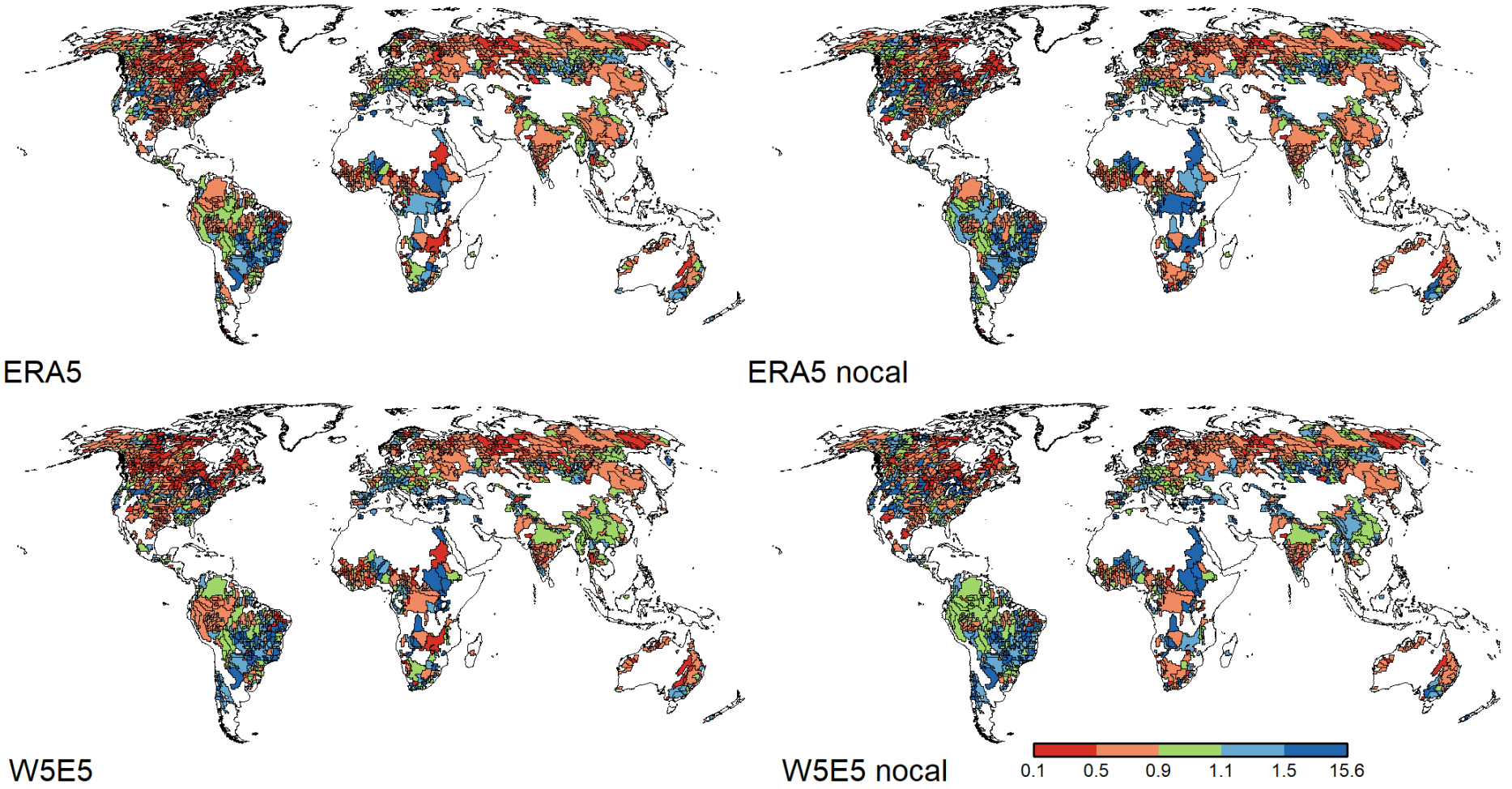

Figure 19: γKGE of modelled discharge calculated for the period between 1979 and 2019 for 1427 basins

Table 12: γKGE values across climate zones for the two uncalibrated model experiments

	γKGE	A	B	C	D	E	Sum
ERA5-nocal	> 1.5	53	19	23	43	4	142
	1.1 - 1.5	66	12	65	76	7	226
	1.1 - 0.9	66	8	75	117	15	281
	0.9 - 0.5	132	48	132	259	8	579
	< 0.5	13	19	12	151	4	199
Sum		330	106	307	646	38	**1427**

	γKGE	A	B	C	D	E	Sum
W5E5-nocal	> 1.5	55	18	32	54	7	166
	1.1 - 1.5	74	28	80	103	16	301
	1.1 - 0.9	85	8	72	92	10	267
	0.9 - 0.5	120	51	94	260	9	534
	< 0.5	9	18	9	112	11	159
Sum		343	123	287	621	53	**1427**

Through calibration of ERA5 the number of basins with γKGE values between 0.9 and 1.1 decreases by 6 %. Although the area covered by basins with good γKGE performance is visually increased and basins form small clusters of good performance such as in central Europe and some Amazon subbasins. Calibration increases performances in climate zone B, C, and E but decreases performance in climate zone A and D.

Calibration of W5E5 decreases the number of basins with good γKGE values by 2 %. Basins with γKGE values between 0.9 and 1.1 form clusters in China, South East Asia, and the Ganges basin. Apart from the clustered areas, the remaining basins with good performance are scattered across the globe with loose concentrations in the northern part of Southern America. Except for climate zone A, calibration increases the performance of W5E5 in all other climate zones.

Table 13: γKGE values across climate zones for the two calibrated model experiments

	γKGE	A	B	C	D	E	Sum
ERA5	> 1.5	50	15	27	26	4	122
	1.1 - 1.5	40	16	64	64	5	189
	1.1 - 0.9	51	14	79	103	16	263
	0.9 - 0.5	157	42	120	277	9	605
	< 0.5	32	19	17	176	4	248
Sum		330	106	307	646	38	**1427**

	γKGE	A	B	C	D	E	Sum
W5E5	> 1.5	55	18	30	28	5	136
	1.1 - 1.5	61	23	78	42	7	211
	1.1 - 0.9	60	18	77	93	14	262
	0.9 - 0.5	143	40	91	294	14	582
	< 0.5	24	24	11	164	13	236
Sum		343	123	287	621	53	**1427**

The performance of both calibrated forcings differs only by one basin, which means that only 18 % of the evaluated basins show good γKGE values. Apart from W5E5's clear superior performance in China and South East Asia, no other significant differences can be identified in figure 19. According to table 13, W5E5 dominates in climate zones A and B, while ERA5 contains slightly more high performing basins in climate zone C, D, and E.

3.2.4 Streamflow indicators

Significant streamflow indicators have been calculated for observed and modelled discharge series. To evaluate the performance of the two climate forcings and the influence of calibration, the deviations between the modelled and observed streamflow indicators have been calculated. The following figures include the calculated streamflow indicator of the observed discharge series in m3 s^{-1} and the deviations of all model experiments from the observed streamflow indicators in percent. Basins with deviations up to 20 % are judged to show good model performance. Basins showing a grey signature indicate an observed streamflow indicator value

of 0, which in turn leads to implausible deviation results created through dividing by 0. Hence, grey signatures provide no information as to whether the model managed to reproduce observed streamflow or not. The same is true for table 14 - 18. The category “basins with deviation greater than 100 %” includes basins with an observed streamflow indicator value of 0 and invalid deviation results as well as those actually deviating greater than 100 %. As a result, the respective category holds little information and will not be discussed further.

Q1

In ERA5-nocal 24 % of all evaluated basins show good compliance of simulated and observed Q1. Basins with up to 20 % deviation are located along the coasts of North America, in the Amazon and Siberian basins (see figure 20 and 21). Basins with deviations greater than 100 % can be seen in the centre of the United States, the east of Brazil and the African and Australian continent. 20 % of W5E5-nocals basins reach good deviation values. Those basins are located along the east coast of the United States, the Amazon basin as well as in China and South East Asia. Basins with a deviation greater than 100 % can be seen in the centre of the United States, the east of Brazil, and the African and Australian continent.

Table 14: Percent deviations of modelled Q1 from observed O1 streamflow

Q1	ERA5-nocal	ERA5	W5E5-nocal	W5E5
0 - 20 %	336	579	291	551
20 - 50 %	416	556	436	554
50 - 100 %	309	202	340	230
> 100 %	366	90	360	92

Through calibration of ERA5 the number of basins with good performance is increased by 72 %. Significant improvements of simulated Q1 are distributed equally across the globe. Major basins with deviations larger than 100 % are the White Nile and Indus. Calibration of W5E5 increases the number of good performing basins by 89 %. Only the Nile and the majority of its tributaries show deviations larger than 100 %.

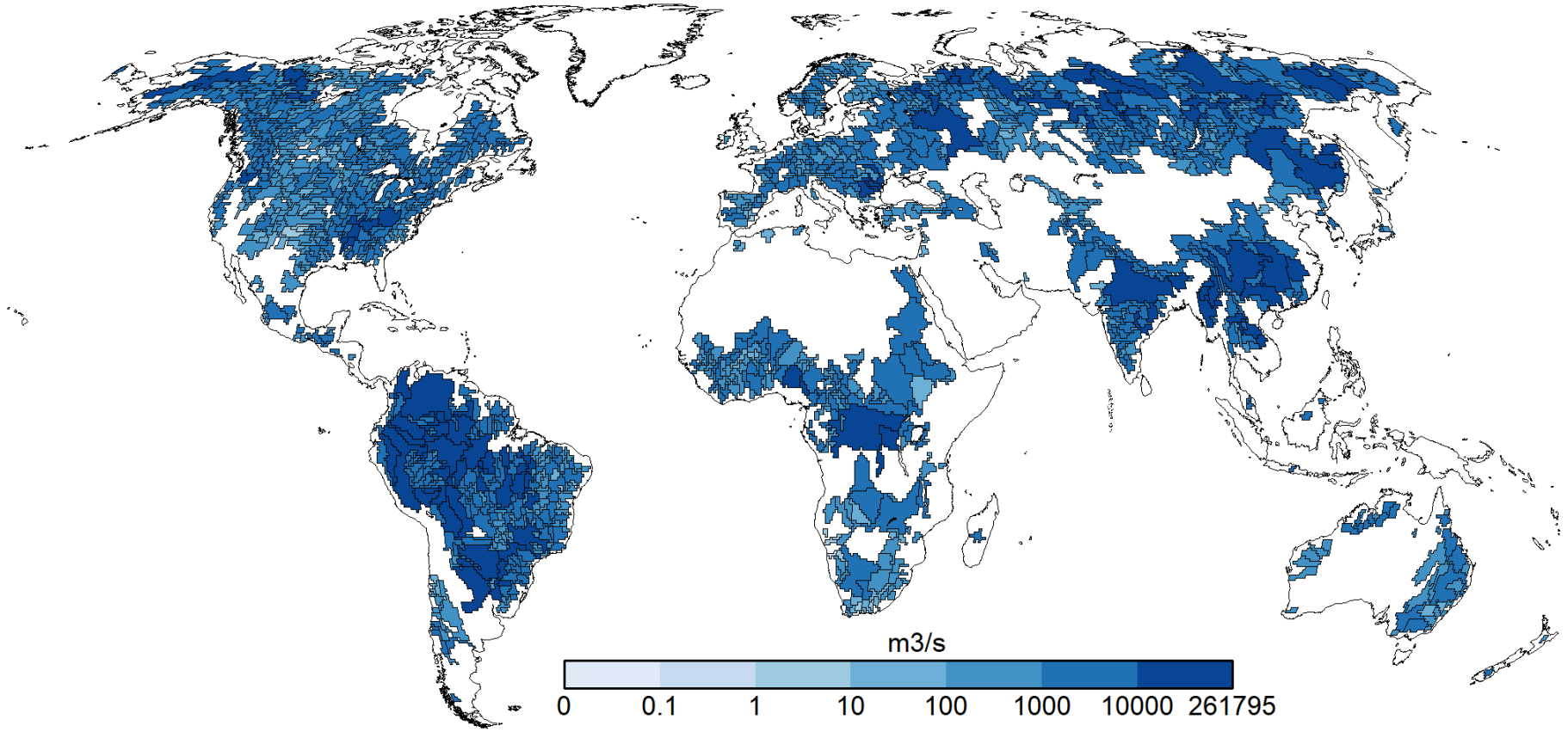

Figure 20: Q1 streamflows ($m^3\ s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

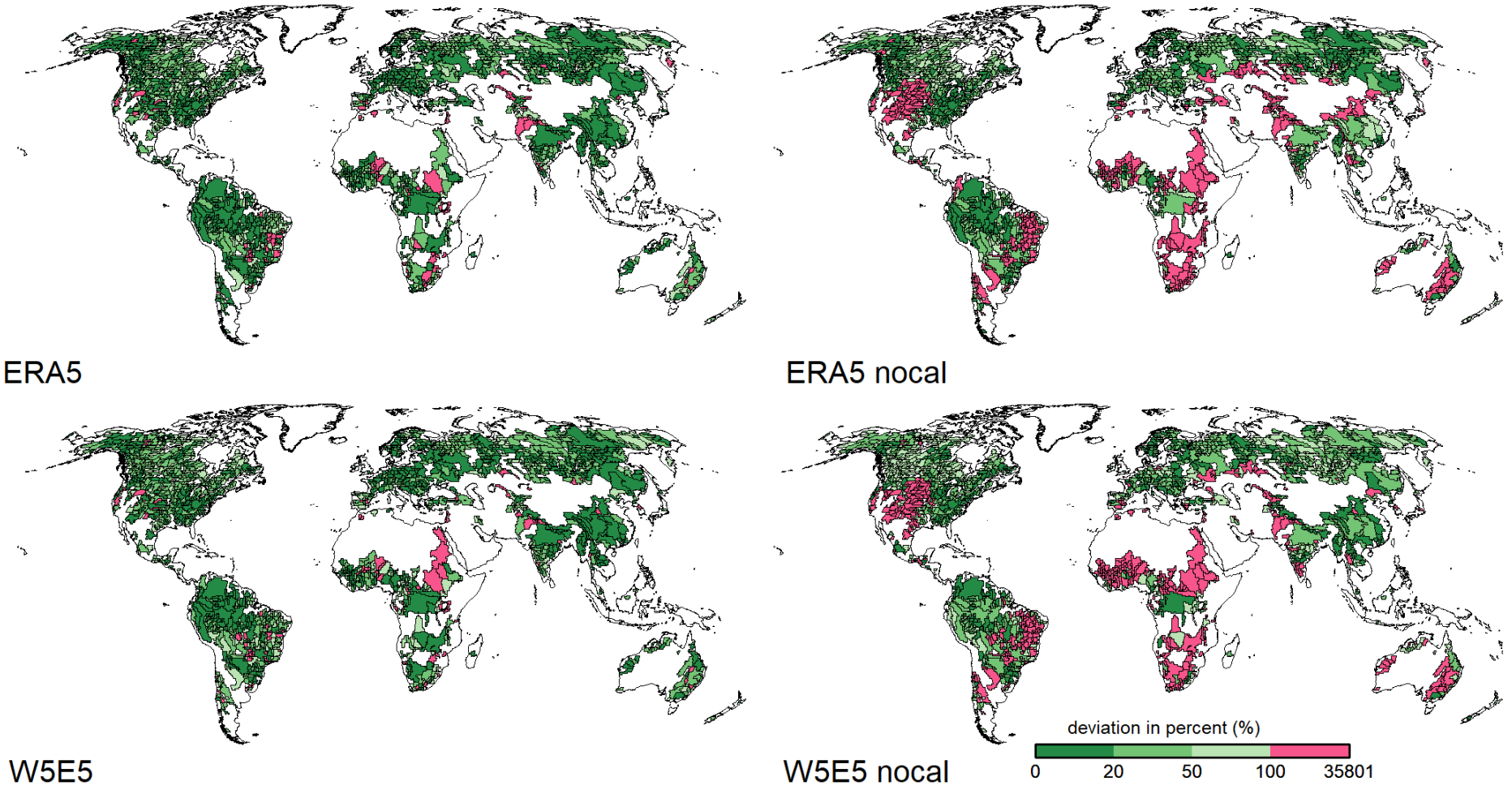

Figure 21: Deviations (%) of modelled O1 flows from observed Q1 flows for 1427 basins

Q10

With ERA5-nocal, approximately 26 % of all basins manage to limit deviation of Q10 to a maximum of 20 %. Basins with low Q10 deviations are distributed along the North American coastlines, the Amazon and south-eastern Siberian basins (see figure 22 and 23). Additionally, they occur incoherently in Europe and Russia. Basins where deviation exceeds 100 % can be found in the centre of the United States as well as on the African continent with the exception of east Africa. Furthermore, almost all Australian and some individual basins in Asia fall below this threshold. W5E5-nocal leads to 23 % of all basins showing good representation of Q10 flows. Significant higher coverage of these basins can be identified on the east coast of North America, the Amazon basin, and western Russia. Basins with deviations greater than 100 % are located in the centre of the United States on the African and Australian continent.

Table 15: Percent deviations of modelled Q10 from observed O10 streamflow

Q10	ERA5-nocal	ERA5	W5E5-nocal	W5E5
0 - 20 %	377	777	326	773
20 - 50 %	431	512	413	510
50 - 100 %	273	100	353	109
> 100 %	346	38	335	35

After calibration of ERA5 the number of basins showing good streamflow compliance is increased by 106 %. These basins are distributed equally over the continents. A concentration of basins with deviations between 20 % and 100 % can be identified surrounding the Ural mountains. Furthermore, the only two larger basins showing deviations greater than 100 % are the Indus and the Australian Cooper Creek. Through calibration of W5E5 the number of basins with good performance is increased by 137 %. Improvements are not limited to a continent or region. Basins surrounding the Ural mountains indicate deviations between 50 and 100 %. Visually basins with deviations greater than 100 % are very spars. The only two larger basins showing such deviations are the Indus tributary Chenab River and the Cooper Creek.

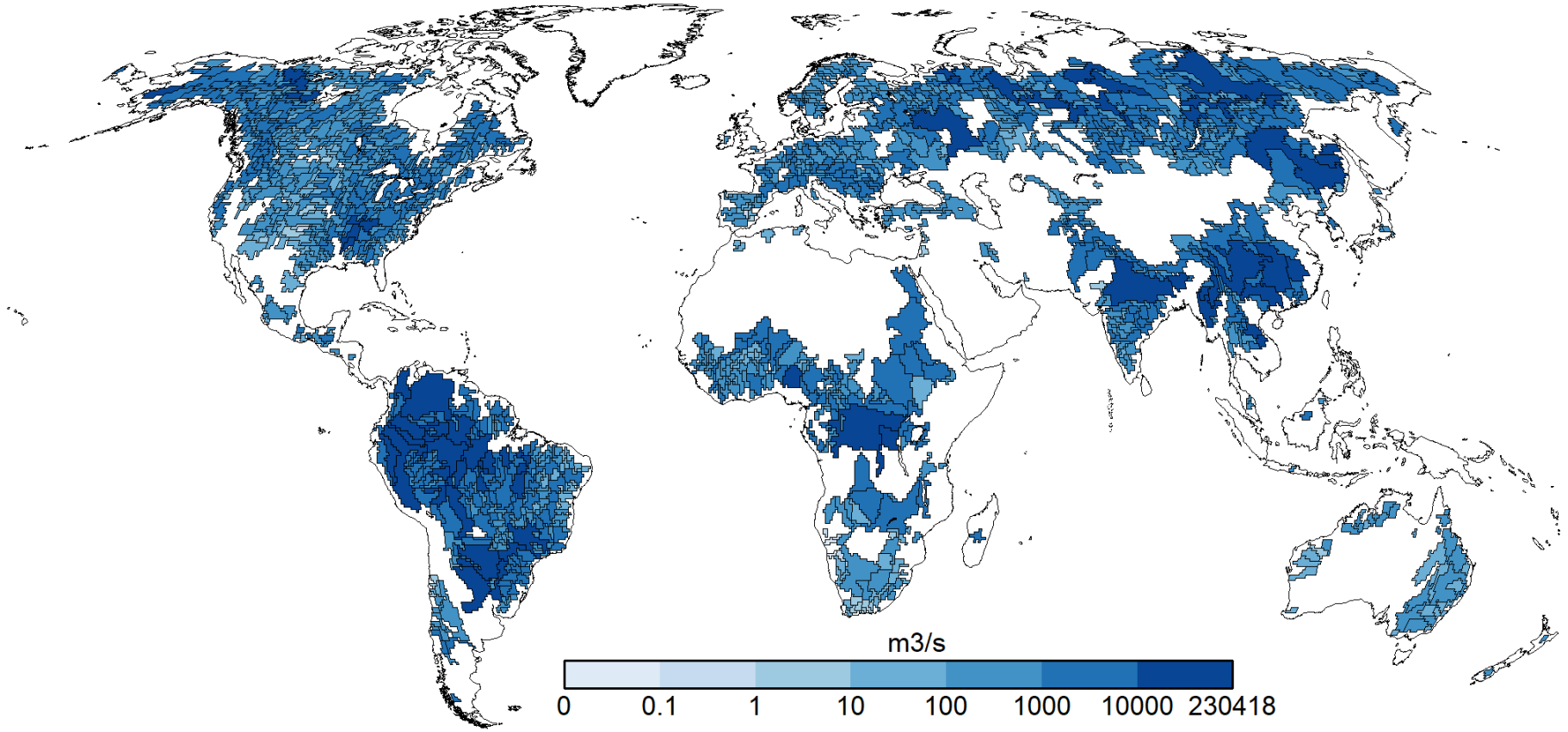

Figure 22: Q10 streamflows ($m^3\ s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

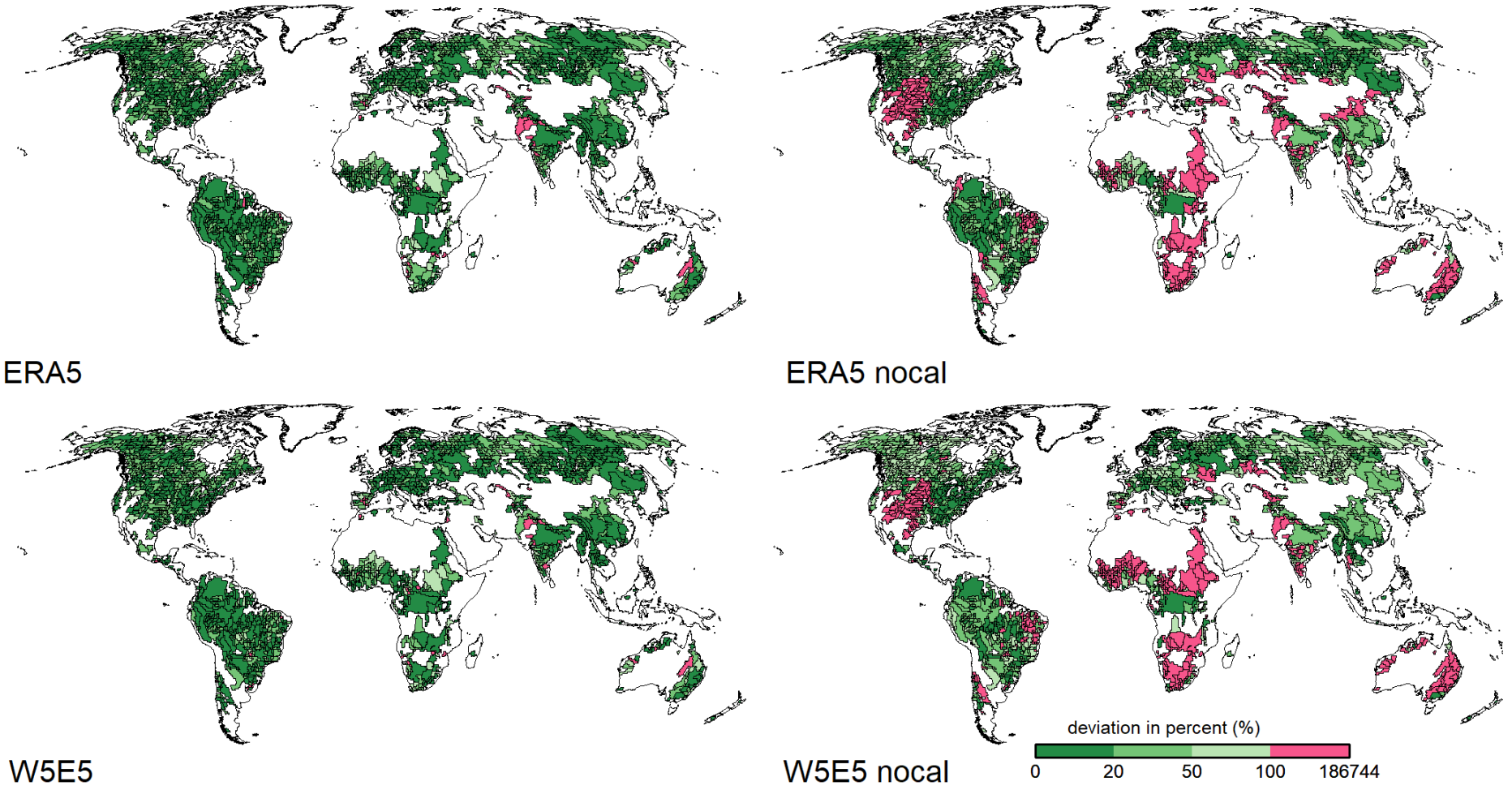

Figure 23: Deviations (%) of modelled O10 flows from observed Q10 flows for 1427 basins

Q50

ERA5-nocal leads to 25 % of all basins reaching good streamflow compliance. Basins with low deviations are located on the coasts of North America and the centre of South America as well as Scandinavia (see figure 24 and 25). Basins with deviations greater than 100 % cover the centre of the United States, most basins on the African and Australian continent, most Indian basins as well as northern Siberian basins. W5E5-nocal leads to 22 % of all basins reaching good streamflow compliance. Those basins are located on the east coast of North America, the south-east of South America, Scandinavia, adjacent Russian basins, and south-eastern Russian basins. Basins with deviations above 20 % but below 100 % are distributed all over. The centre of the US, most basins on the African and Australian continent as well as most Indian basins and north eastern Siberian basins are dominated by deviations above 100 %.

Table 16: Percent deviations of modelled Q50 from observed O50 streamflow

Q50	ERA5-nocal	ERA5	W5E5-nocal	W5E5
0 - 20 %	352	597	311	603
20 - 50 %	370	354	383	335
50 - 100 %	260	224	328	240
> 100 %	445	252	405	249

Calibration increases the number of basins with deviations equal to or below 20 % by 144 % between ERA5-nocal and ERA5. Those basins are concentrated in North and South America but also in Europe and the western part of Russia as well as in South East Asia. Most African basins previously showing deviations above 100 % now improved to deviations below 100 %. The same is true for the Indus and Ganges as well as Chinese basins. W5E5 shows an increase of 107 % in the number of basins with deviations below or equal to 20 %. Those basins are concentrated in North and South America as well as in Europe, Scandinavia and western Russian basins, China and South East Asia. As for ERA5, most African basins now show deviations between 20 and 100 %. Basins with deviations greater than 100 % are largely limited to Russian high latitudes but also occur in India and Australia.

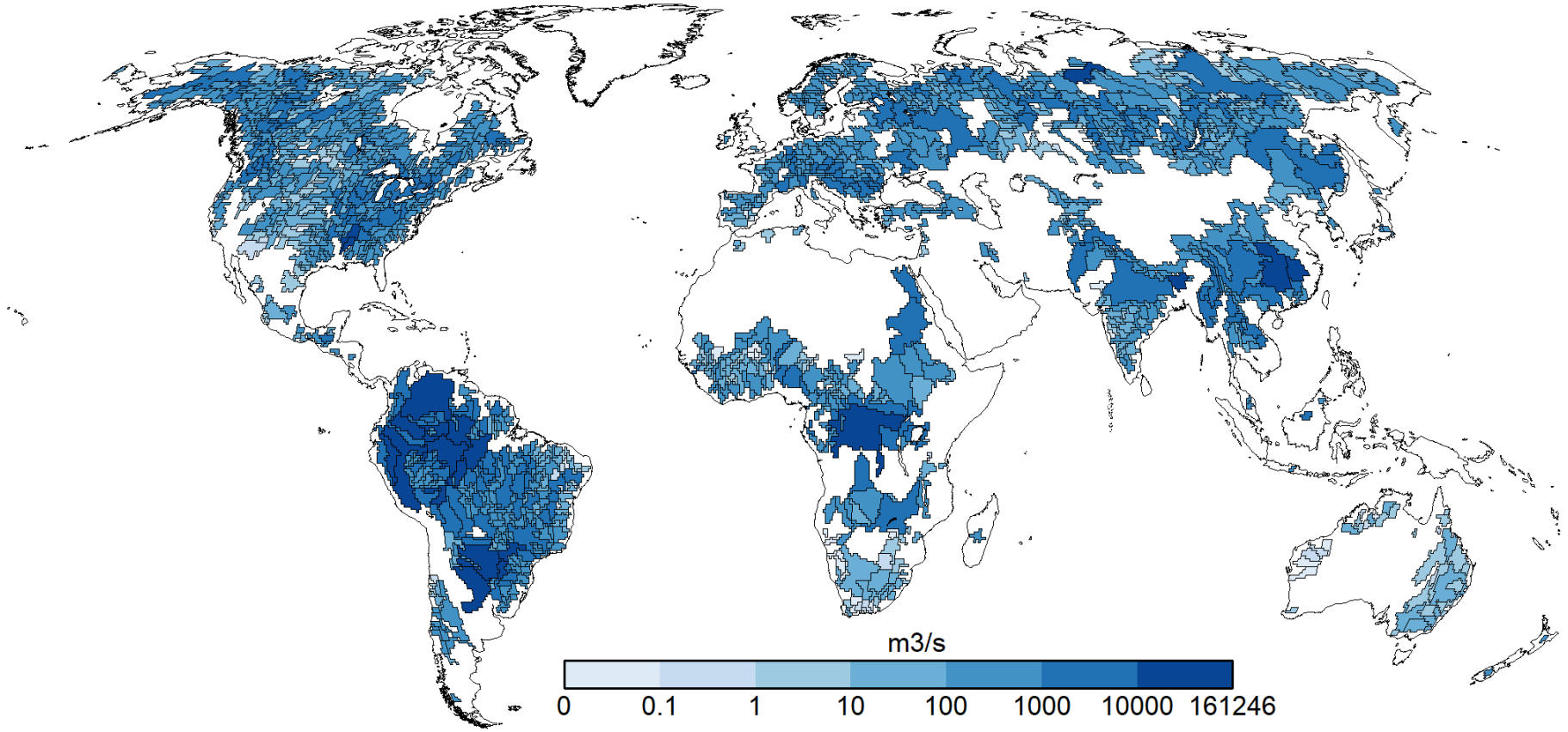

Figure 24: Q50 streamflows ($m^3\ s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

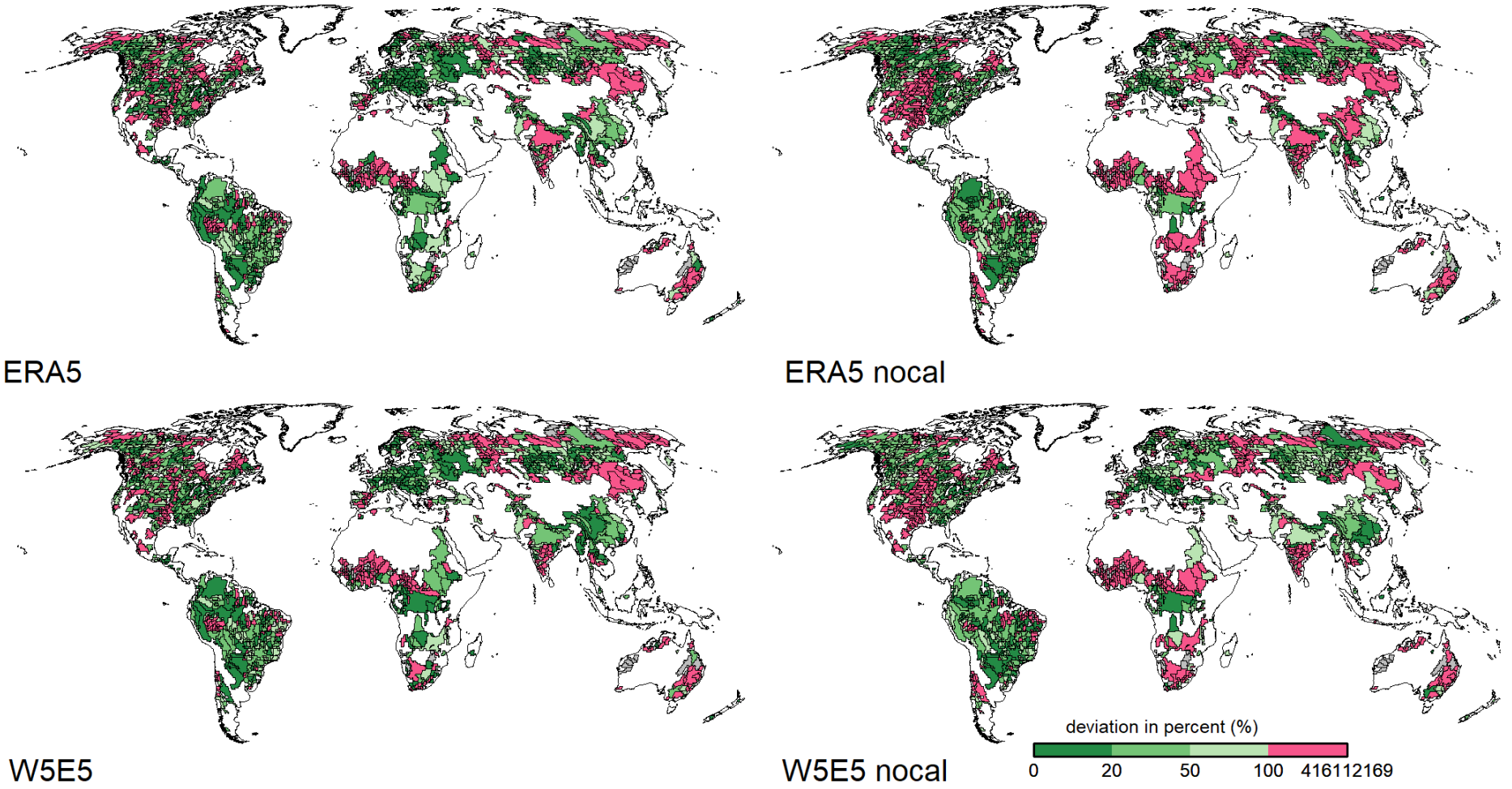

Figure 25: Deviations (%) of modelled O50 flows from observed Q50 flows for 1427 basins

Q90

ERA5-nocal results in 17 % of all basins reaching good streamflow compliance. Only loose concentrations of these basins can be identified (see figure 26 and 27). Basins with deviations larger than 100 % are located in the centre of North America, India, Australia and are widely distributed over the African continent and Russia. W5E5-nocal manages to reach a good representation of Q90 in 15 % of all evaluated basins. They are distributed over all continents. However, they rarely form cluster such as in the Argentinian Rio Parana and adjacent basins as well as the Danube basin and lower reach Yangtze. Deviations greater than 100 % can be seen in basins in the centre of North America, India, Australia and on the African continent as well as surrounding the Ural mountains and the eastern Russia.

Table 17: Percent deviations of modelled Q90 from observed O90 streamflow

Q90	ERA5-nocal	ERA5	W5E5-nocal	W5E5
0 - 20 %	246	286	220	313
20 - 50 %	280	308	293	311
50 - 100 %	336	377	392	344
> 100 %	565	456	522	459

After calibration of ERA5 the number of good performing basins is increased by 16 %. Calibration leads to the formation of low-deviation cluster in central Europe and western Russia as well as central Siberia. The distribution of basins with deviations greater than 100 % is reduced but can still be seen in India, eastern Russia, and Sub-Sahara Africa. Calibration of W5E5 leads to an increase of 42 % in basins reaching good streamflow compliance. Some cluster already existing in W5E5-nocal have expanded in W5E5, e.g. the Rio Parana region. But most clusters are exclusive for W5E5. Those are located in the Amazon basin, eastern Europe and western Russia as well as Siberia and upper reach Yangtze including adjacent basins. Basins with deviations greater than 100 % are located in Sub-Saharan Africa, Australia and eastern Russia as well as surrounding the Ural mountains.

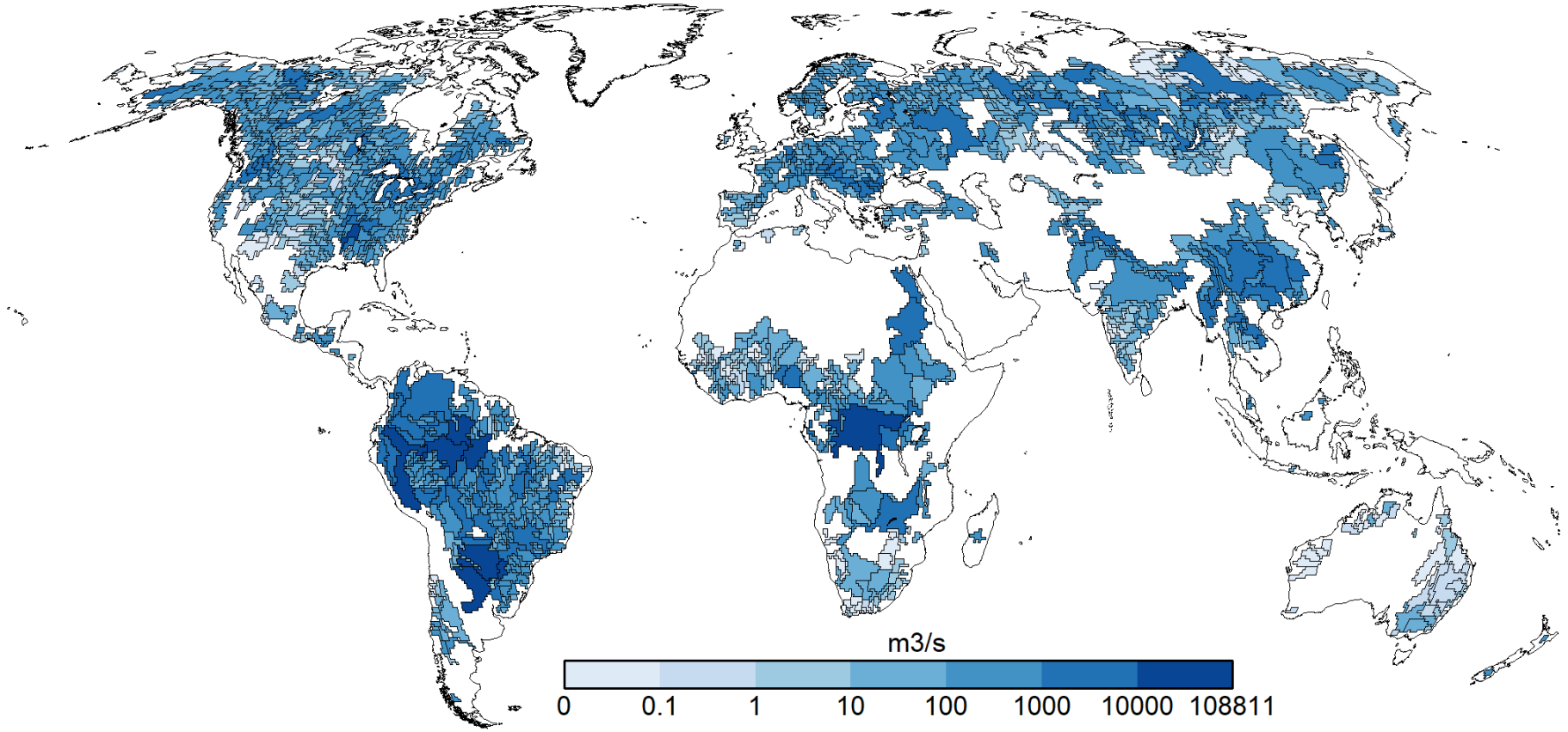

Figure 26: Q90 streamflows ($m^3\ s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

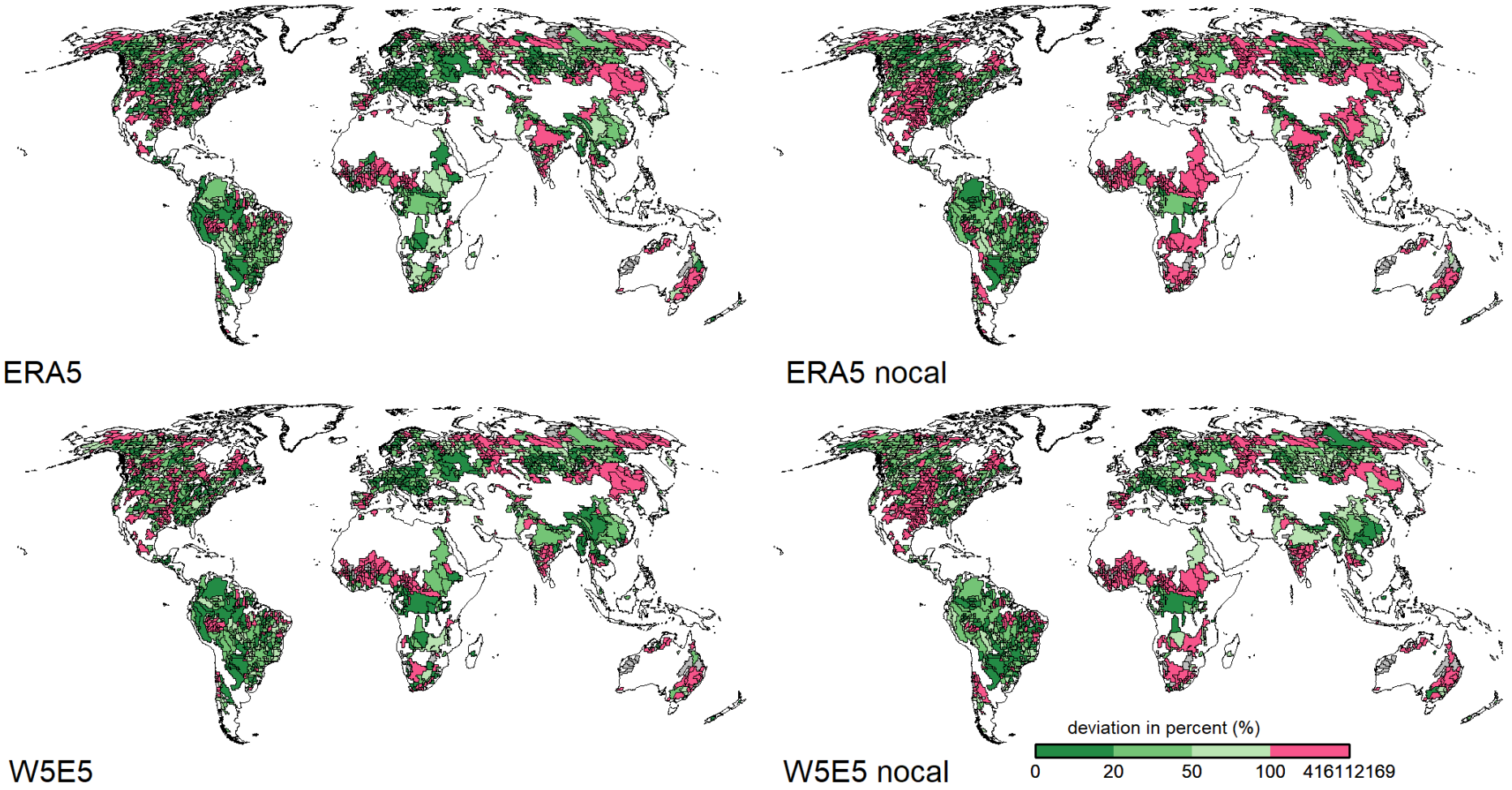

Figure 27: Deviations (%) of modelled O90 flows from observed Q90 flows for 1427 basins

Q99

16 % of all evaluated basins show good streamflow compliance for ERA5-nocal. Those basins are located in western Canada, the Amazon and Rio Parana basins as well as central Europe and Siberia (see figure 28 and 29). Basins with deviations greater than 100 % are concentrated in the centre of North America, India, the north-wester China as well as eastern Russia and surrounding the Ural mountains. The African continent is dominated by basins with deviations above 100 %. W5E5-nocal contains 186 basins (13 % of all basins) with good streamflow compliance. Those basins are distributed all over the continents, and no clear concentration can be identified. Basins with deviations greater than 100 % are concentrated in the centre of North America and occur quite frequently on the African continent as well as in Russia.

Table 18: Percent deviations of modelled Q99 from observed O99 streamflow

Q99	ERA5-nocal	ERA5	W5E5-nocal	W5E5
0 - 20 %	230	227	186	223
20 - 50 %	253	260	268	293
50 - 100 %	365	433	458	419
> 100 %	579	507	515	492

Calibration shows close to no effect on the number of well-performing basins in ERA5 (differences: 3 basins). However, their spatial distribution changes. A small cluster can be seen on the Canadian-Alaskan boarder but also basins in Europe and western Russia show a decrease of deviation. The most significant difference can be seen on the African continent, where a considerable amount of basins previously showing deviations greater than 100 % now show deviations either between 20 and 100 % or even below 20 %. The same is true for basins in China and South East Asia. Calibration of W5E5 increases the number of basins with good streamflow compliance by 20 %. These basins are concentrated in China and South East Asia, Siberia and western Russia. Additionally, the distinct pattern of basins with deviations greater than 100 % in central North America has dissolved, and a more heterogeneous pattern can be identified. Concentrations of basins with deviations greater than 100 % concentrate around the Ural mountains and eastern Russia.

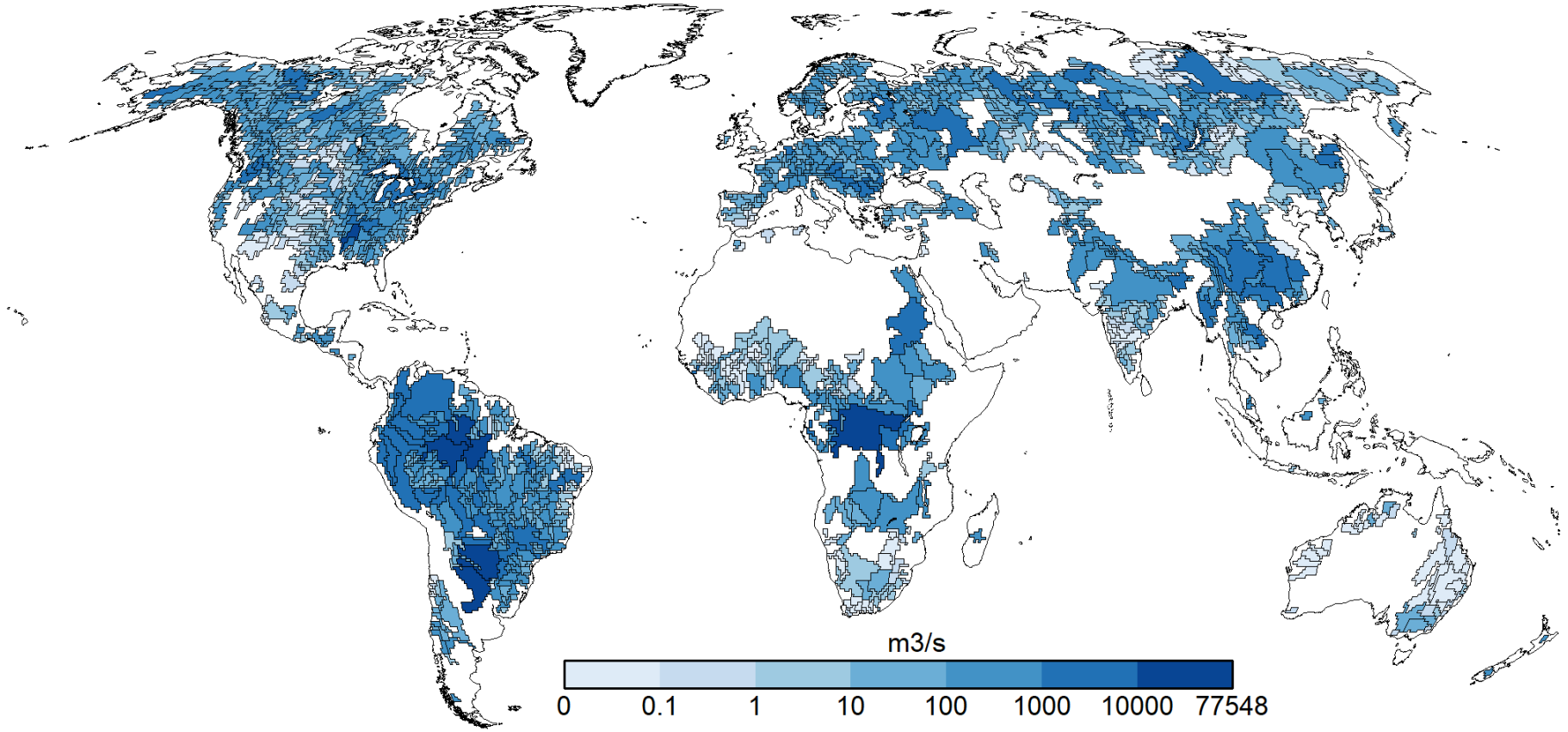

Figure 28: Q99 streamflows ($m^3 s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

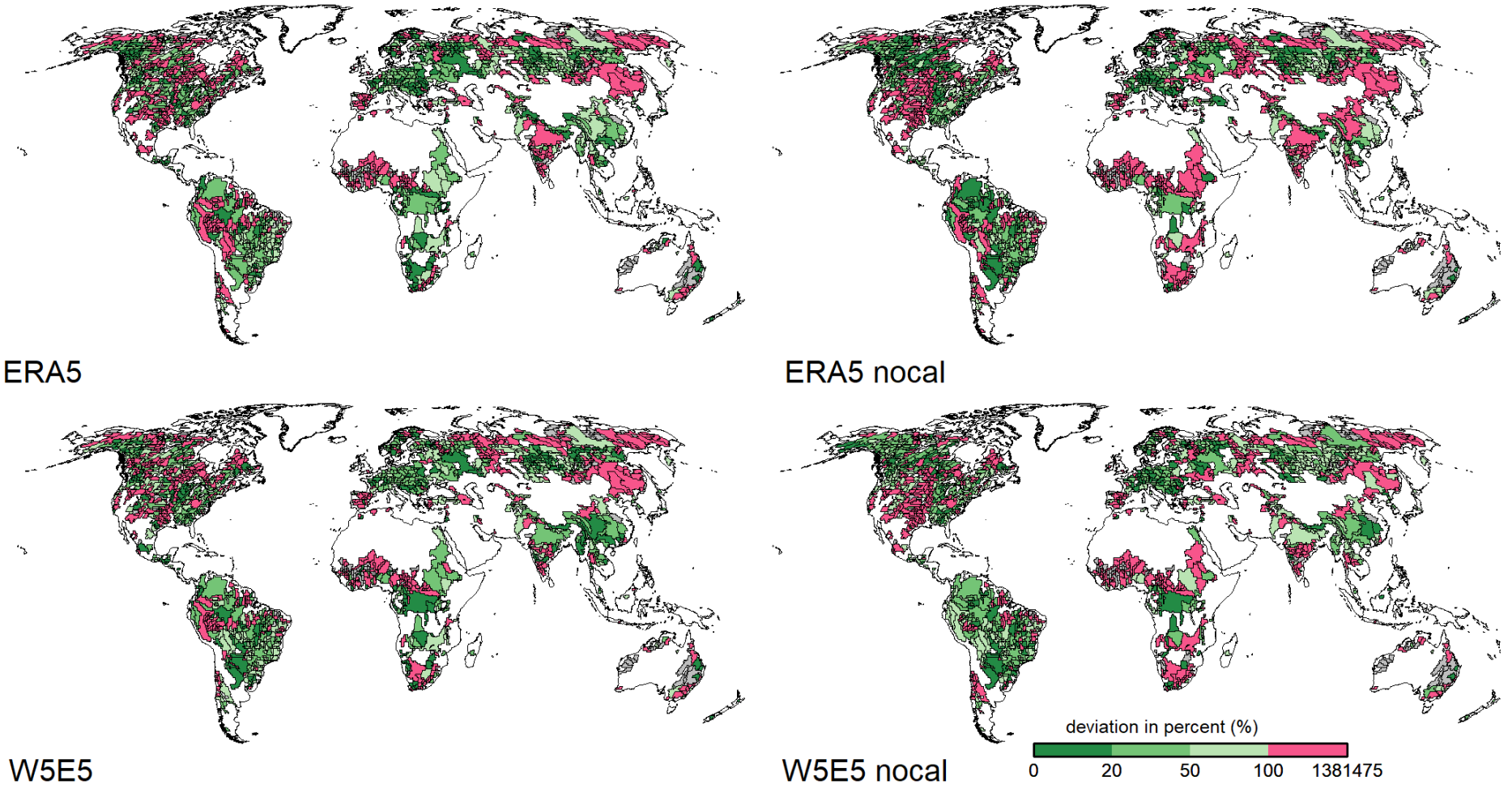

Figure 29: Deviations (%) of modelled O99 flows from observed Q99 flows for 1427 basins

3.2.5 TWSA

R2 - CSR-RL06

R^2 describes the proportion of the variance of observed values that the model can reproduce. One is again the value aimed for. Zero, on the contrary, indicates no correlation between the model and observed values. However, values equal to or larger than 0.8 can already be judged as good model performance. 43 of the evaluated basins show R^2 values above 0.8 in ERA5-nocal. Basins with optimal performance are concentrated in the northern part of South America and Russian high latitude basins (see figure 36). ERA5-nocal TWSA simulations yield low R^2 values for basins in the centre of North America, south and east Africa. One basin located in Siberia shows an R^2 value below 0.2. W5E5-nocal includes three basins with a R^2 value above 0.8 less than ERA5-nocal. High-performing basins are concentrated in South America, China, and South East Asia as well as basins in northwest Russia. Additionally, single high-performing basins can be identified in North America and in Africa. Only one basin in Siberia shows a R^2 value below 0.2. Other low-performing ($0.2 < R^2 < 0.6$) basins can be identified in the centre of North America, South Africa, and in eastern Russia

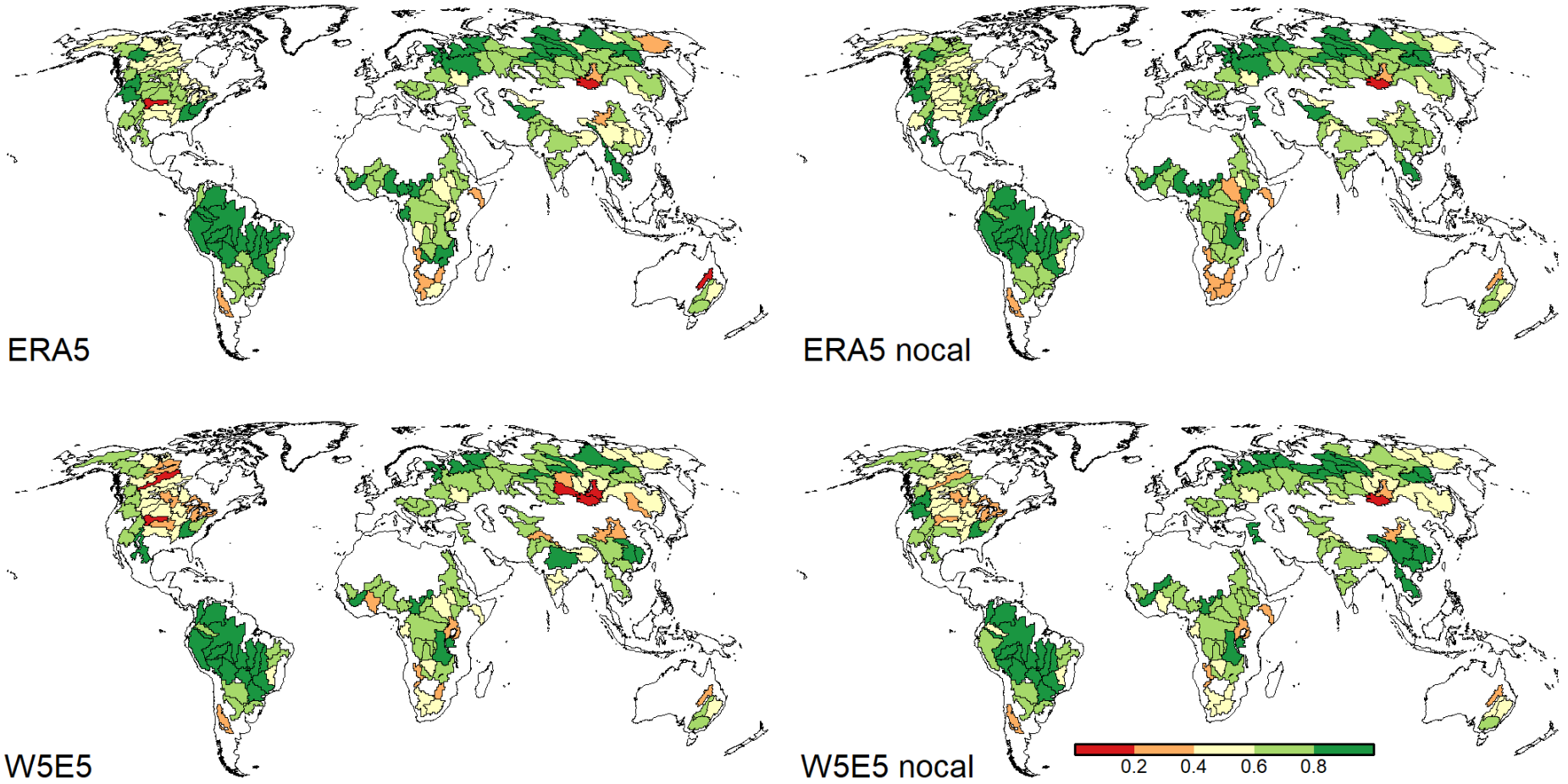

Figure 30: Coefficient of determination for TWSA of all four model experiments evaluated with CSR-RL06

Through calibration of ERA5 the number of basins with optimal values is reduced by two. The number of basins with R^2 values below 0.2 increases by two. However, performance over the northern part of South America, central North America as well as Africa increases while it slightly decreases for Russian and Asian basins. Calibration of W5E5 reduces the number of basins with R^2 values above 0.8 by approximately 25 %. While performance increases in South America, it decreases over China, South East Asia, and Russia. Basins that fell into the classes of low-performance in W5E5-nocal, have either not changed or transitioned to a lower class. The number of basins with R^2 below 0.2 has increased to five, located either in Siberia or in the centre of North America.

W5E5 has 27 % less high-performing basins than ERA5. Apart from South America, where both forcings show coherently high performing basins, ERA5 shows higher R^2 values on all other continents. Additionally, W5E5 includes more basins with R^2 values close to zero.

R^2 – JPL-RL06M

ERA5-nocal reaches high R^2 values in 36 basins. High-performing basins are located in South America. Individual high-performing basins can be found in Asia, North America and Africa (see figure 31). Low-performing basins ($0.2 < R^2 < 0.6$) are clustered in central North America but can also be found on the African continent and in eastern Russia and with less frequency in Asia. Three basins (White Nile, Yukon and Selenga) fall below the mark of 0.2. W5E5-nocal shows 37 high performing basins, of which the majority are coherently located in South America, China and South East Asia. Individual high-performing basins are located in North America, Sub-Saharan Africa and western Russia. Low-performing basins are distributed in Alaska, central North America and eastern Russia as well as central to southern Africa. In total, three basins (Slave River, Selenga and its tributary) fall below a R^2 value of 0.2.

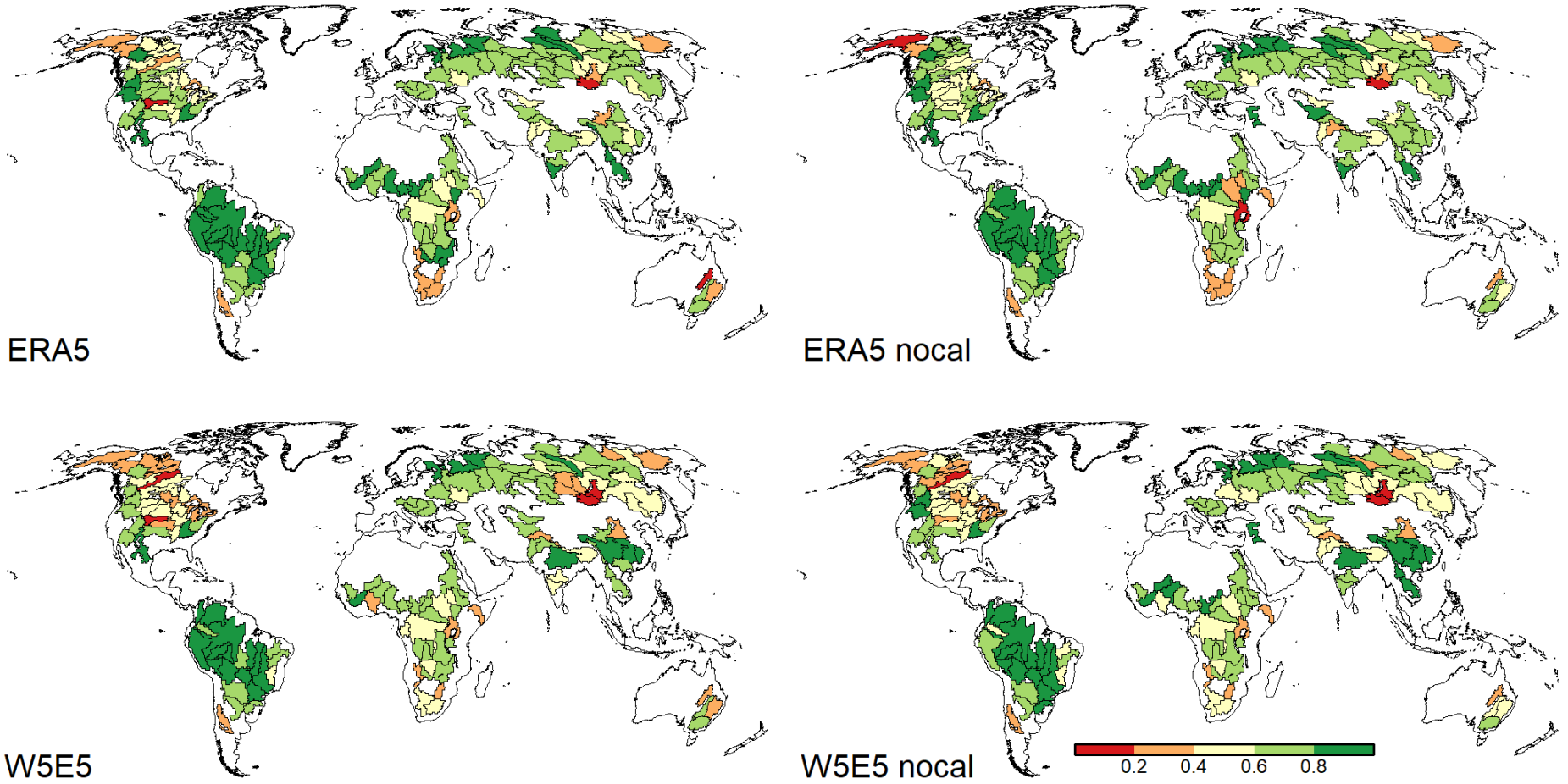

Figure 31: Coefficient of determination for TWSA of all four model experiments evaluated with JPL-RL06M

Calibration of ERA5 marginally changes the number of basins per R^2 class. In fact, the number of basins in the highest class remains to be 36 and the same is true for the lowest class. However, the spatial distribution of the basins falling into these classes changes. Performance in South America is further increased. Slight performance increases can also be seen over the African continent. Performance over North America is highly heterogeneous since basins in the centre improved to R^2 values between 0.6 and 0.8, but at the same time, some basins previously reaching this class dropped down to values below 0.6. The Yukon basin however improved performance and now shows a R^2 value between 0.2 and 0.4. On the contrary, the Platte River basin in the centre of the US fell below the 0.2 mark. Additionally, Australian Cooper Creek basin shows an R^2 value below 0.2. Through calibration, the number of basins with R^2 larger than 0.8 is reduced by 32 % in W5E5. However, the spatial distribution of high-performing basins remains largely the same. The spatial pattern of low-performing basins is unchanged as well. The number of basins with a R^2 value below 0.2 is increased by one basin (Platte River basin).

ERA5 contains eleven more basins with R^2 values higher than 0.8 compared to W5E5. Both forcings perform well in South America, but W5E5 outperforms ERA5 in China and South East Asia. ERA5 shows superior performance in North American, African and Russian basins. Both forcings show the lowest R^2 values in basins located in central North America however, W5E5s performance is significantly weaker.

bR² – CSR-RL06

ERA5-nocal includes 14 basins with high performance values equal to or above 0.8. The highest density of high-performing basins can be seen in South America (see figure 32). Performance over North America shows a diverse pattern. While basins closer to either of both coasts perform medium ($0.6 < b R^2 < 0.8$) to good ($b R^2 < 0.8$), basins located in the centre show low performances. Africa, Asia and Australia are largely dominated by basins with $b R^2$ values below 0.4. In total, 25 basins show $b R^2$ values below 0.2. W5E5-nocal includes just nine basins with good performance. Those basins are distributed across the continents, with the exception of Africa and Australia. 17 % of all evaluated basins fail to surpass the $b R^2$ value of 0.2. A clear concentration of these basins can be identified in the centre of North America.

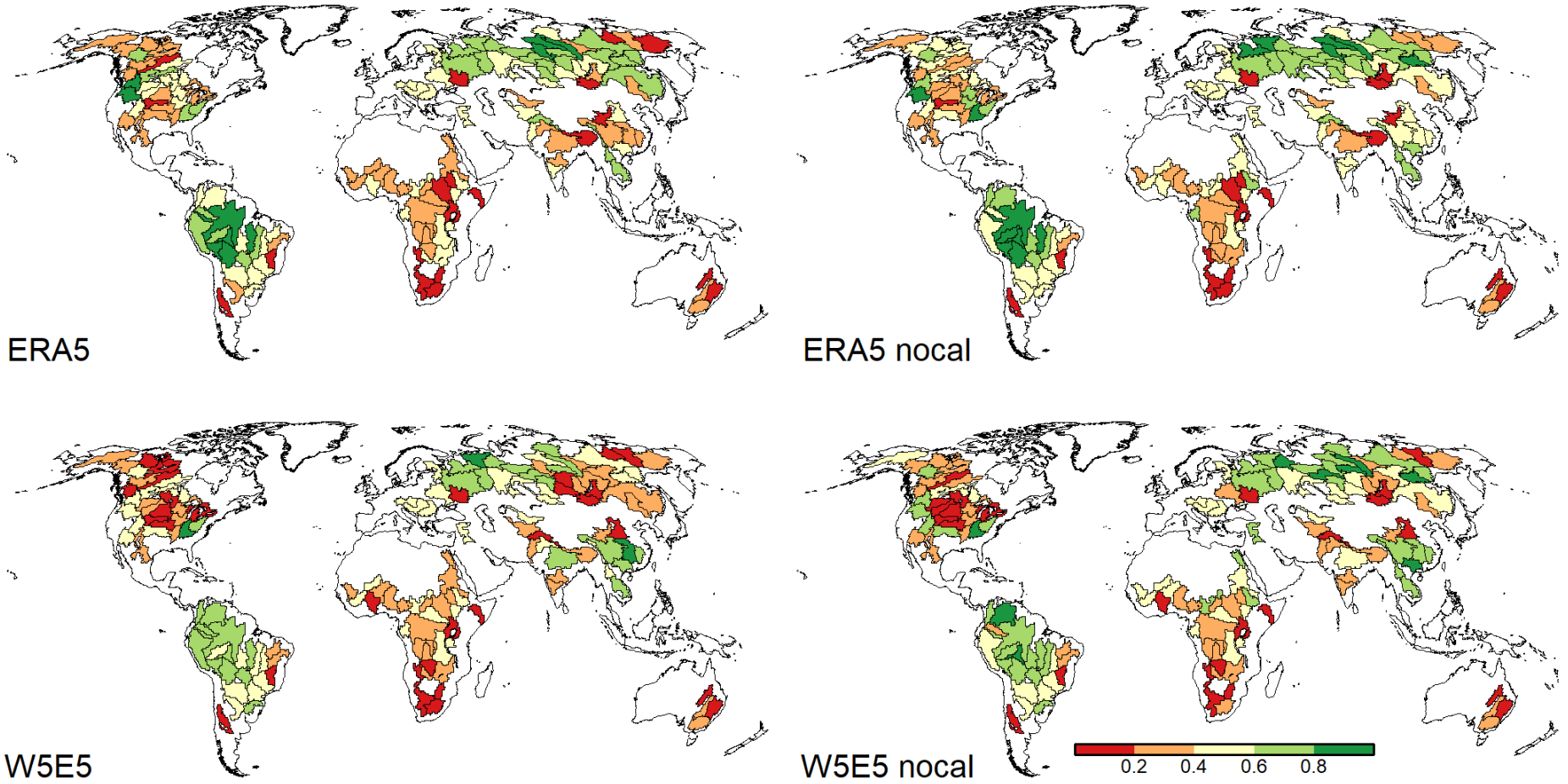

Figure 32: bR^2 for TWSA of all four model experiments evaluated with CSR-RL06

Calibration of ERA5 reduces the number of high-performing basins, which can be seen equally across the continents. Almost half of all evaluated basins show *b*R² values below 0.4. Calibration of W5E5, on the other hand, reduces the number of high-performing basins to just three (Yangtze, Ohio River and Pechora). Simultaneously the number of basins with bR² values below 0.2 is increased by 20 %. The spatial concentration of these basins in North America is persistent. ERA5 contains more than three times as many high-performing basins and outperforms W5E5 everywhere apart from China and South East Asia.

bR^2 – JPL-RL06M

Only 6 % of ERA5-nocal basins show *b*R² values above 0.8. They are located in the Amazon basin, the west coast of North America or in higher Russian latitudes (see figure 33). While most continents show a considerable amount of medium-performing basins (0.6< *b*R² <0.8), African basins result dominantly in values below 0.4. 16 % of all evaluated basins fail to reach *b*R² values above 0.2. The equal distribution of those basins prohibits the identification of regions with particular low performance. Comparatively, W5E5-nocal leads to 4 % of all basins showing high *b*R² values. They are located in South America, South East Asia, China, and northern Russia. On the contrary, 24 % of evaluated basins fail to reach *b*R² values above 0.2. Basins falling into the lowest performance class are distributed all over the globe. However, a significant concentration can be identified in North America.

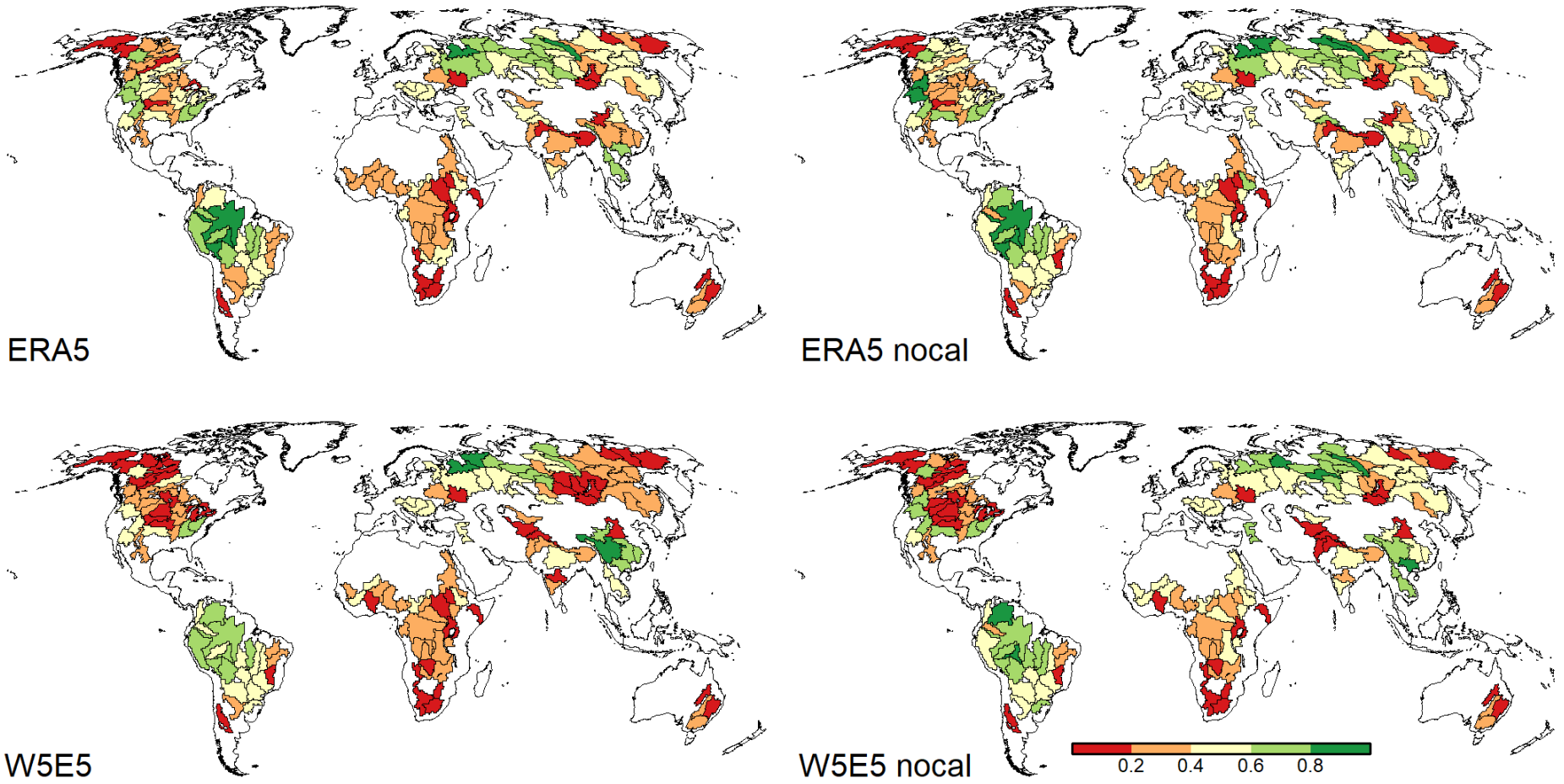

Figure 33: *b*R^2 for TWSA of all four model experiments evaluated with JPL-RL06M

Calibration of ERA5 reduces the number of high-performing basins to five, located in the Amazon basin and high Russian latitudes. Performance in all other regions is reduced considerably. 32 % of all basins show $b\mathrm{R}^2$ values between 0.2 and 0.4, making it the most dominant class for ERA5. W5E5 includes just three high-performing basins. All of which were not included in this class before calibration. Regions with high or medium performing basins are the Amazon basin, northeastern Russian as well as Chinese and South East Asian basins. The number of basins reaching $b\mathrm{R}^2$ values below 0.2 increases and comprises 27 % of all evaluated basins. The concentration of basins in the lowest class is persistent in North America. Another cluster can be identified surrounding Lake Baikal.

ERA5 includes two more high-performing basins than W5E5. W5E5 shows higher performance in China and South East Asia, while ERA5 dominates in South America and north-eastern Russian basins. Both forcings perform inferior on the African, North American and Australian continents. W5E5 contains 15 more basins with $b\mathrm{R}^2$ values below 0.2 than ERA5.

γKGE – CSR-RL06

ERA5-nocal and W5E5-nocal each contain four basins with high performance ($0.9 < \gamma\mathrm{KGE} < 1.1$). However, the model seriously under- or overestimates variability over most land areas when run with either forcing (see figure 34). Regions where ERA5-nocal only marginally over- or underestimates variability ($0.5 < \gamma\mathrm{KGE} < 0.9$ or $1.1 < \gamma\mathrm{KGE} < 1.5$) are the Amazon basin and China. The region where W5E5-nocal only slightly deviates from the optimum value is limited to basins in China. Compared to ERA5-nocal, W5E5-nocal tends to underestimate variability in this region. Additionally, W5E5-nocal forms more distinct cluster of either over- or underestimation of variability. For example, Amazon tributaries and western Russian basins uniformly underestimate variability while eastern Russian and Canadian basins tend to overestimate variability.

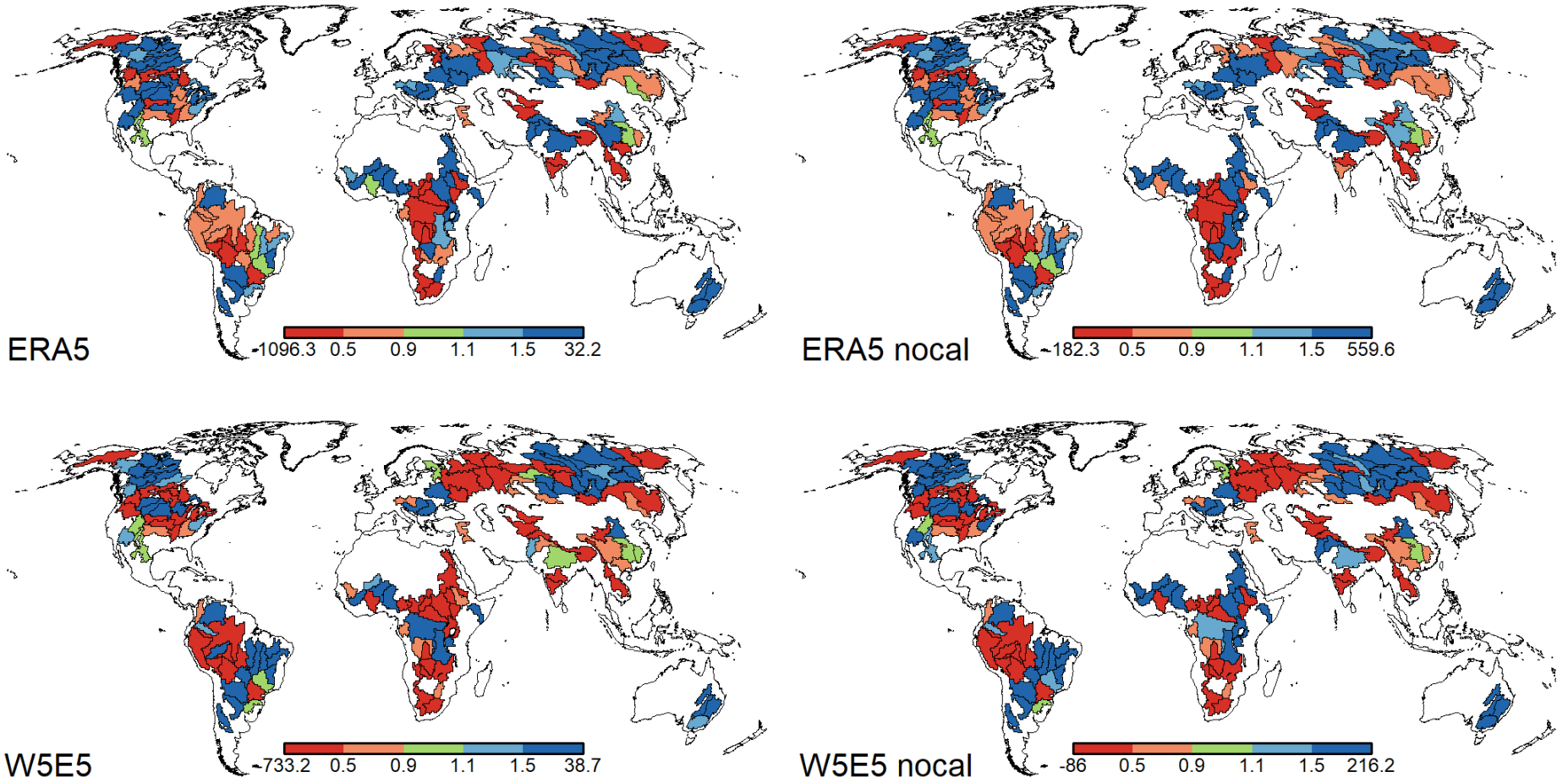

Figure 34: γKGE for TWSA of all four model experiments evaluated with CSR-RL06

The number of good performing basins is doubled through calibration between ERA5-nocal and ERA5.Still, it is not possible to identify a region with coherently high-performing basins as they are scattered across the continents. As for ERA5-nocal, the Amazon basin performs relatively well. However, performance of Chinese basins is slightly reduced. Additionally, ERA5 tends to significantly overestimate variability in Siberian basins. Calibration of W5E5 results in twice as many high-performing basins. The clustering of seriously under- and overestimated variability is persistent in W5E5 and Chinese basins remain the relatively best performing ones. Calibration leads to an inversion of the variability signal from seriously over- in W5E5-nocal to seriously underestimating variability in W5E5 in the Nile basin.

Both forcings are improved through calibration and show almost the same number of high-performing basins (difference: 1 basin). Furthermore, both forcings overestimate variability in eastern Russian and central and Canadian, central US and east Brazilian basins. The variability signal of the Nile is reversed between the forcings. While W5E5 seriously underestimates variability, ERA5 overestimates variability.

γKGE – JPL-RL06M

ERA5-nocal leads to eight high-performing basins, of which half are located in South America (see figure 35). W5E5-nocal contains only four high-performing basins with no apparent concentration in any region. Both forcings seriously over- or underestimate variability for most basins. However, ERA5-nocal shows a relatively good performance in the Amazon basin, where variability either is in line with the optimum values or is slightly underestimated. W5E5-nocal shows good representation or slight overestimation of variability in basins in China. Both forcings show cluster of seriously overestimated variability in east Russian, Canadian, central US and east Canadian basins.

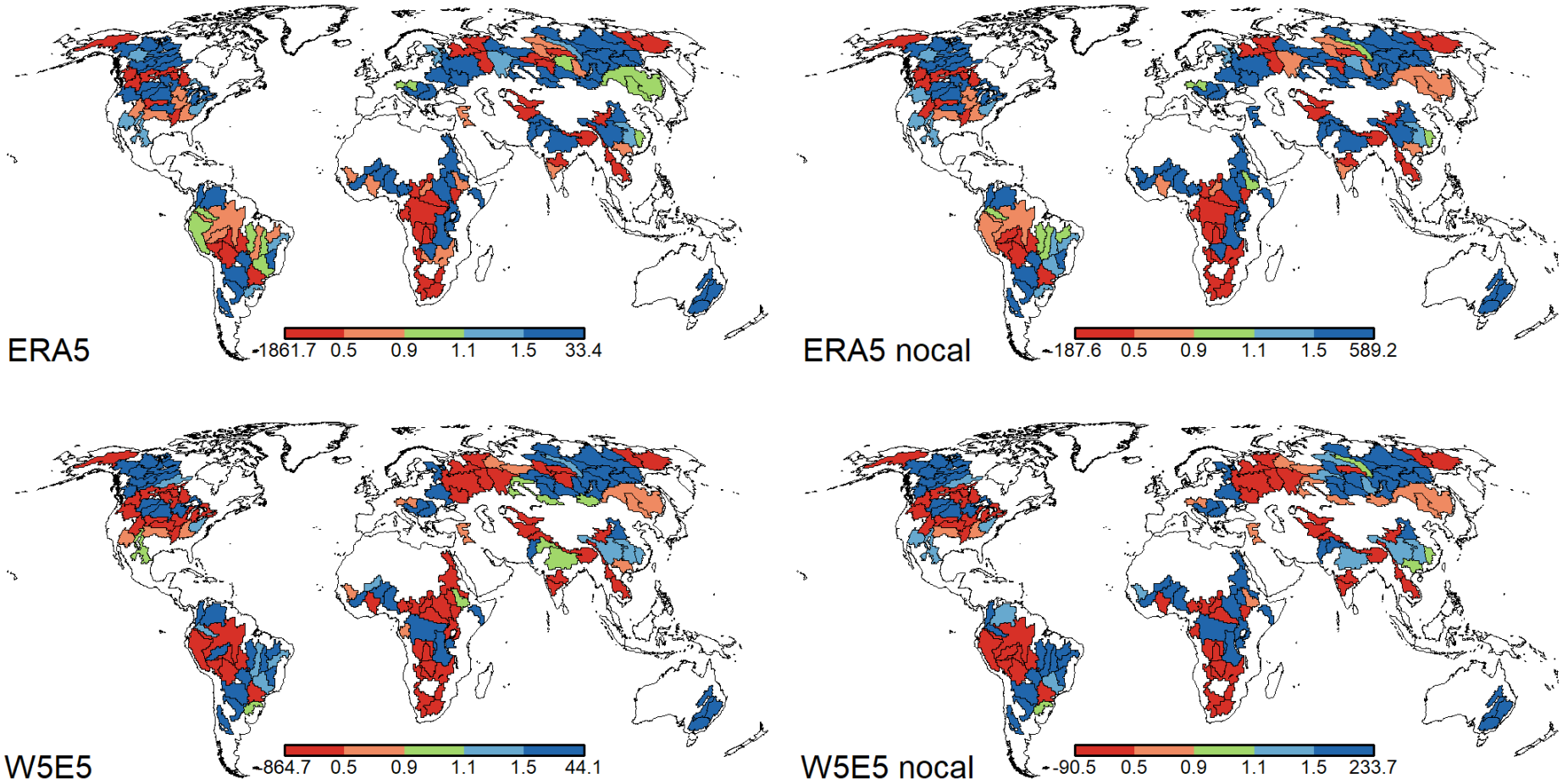

Figure 35: γKGE for TWSA of all four model experiments evaluated with JPL-RL06M

Calibration slightly increases the performance of ERA5. Two regions with relatively good performance can be identified: the Amazon basin and basins along the Russian and Chinese border. Apart from that, the spatial distribution of seriously over- and underestimation of variability remains the same between ERA5-nocal and ERA5. Through calibration of W5E5, the number of high-performing basins is almost doubled. Chinese basins still perform relatively well even though calibration negatively affected them. Ganges and one Indus tributary form a small cluster of high performance. Apart from that, high performing basins appear only individually. The cluster of over- and underestimation identified in W5E5-nocal persists in W5E5. The variability signal is reversed in the Nile basin and shows an underestimation of variability in W5E5.

ERA5 contains three high-performing basins more than W5E5. ERA5 performs superior in South America, while W5E5 slightly outperforms ERA5 in China. Nevertheless, over- or underestimating variability is identified for most basins in both forcings. Both overestimate variability in east Russian, Canadian, central US and east Canadian basins. The variability signal is reversed in Congo and Nile basins.

TWSA Trends of JPL-RL06M and CSR-RL06

JPL-RL06M and CSR-RL06 agree in TWSA trends in the majority of the evaluated basins (see figure 36 and 37). Both solutions show positive TWSA trends of 1 to 10 mm yr^{-1} in most central North American basins. Around the Great Lakes, an even higher positive trend of up to 20 mm yr^{-1} can be identified. Comparatively, all North American basins in high latitudes show decreasing TWSA trends. JPL-RL06M shows a greater amplitude of negative trend values than CSR-RL06. A more uniform decreasing TWSA trend can be seen for southern North American basins. Apart from one tributary (Rio Xingu) the entire Amazon basin shows an increasing trend of TWSA. Basins in the east of South America show low to strong decreasing trends (-1 to -30 mm yr^{-1}). The majority of African basins show increasing TWSA trends between 1 and 10 mm yr^{-1}. However, basins in South Africa show no long-term decreasing or increasing trend at all. While the European and western Russian basins show a decreasing trend, basins east of the Ural mountains show rather increasing or stable TWSA trends. These conditions are disrupted on the transition between central and east Siberia, where basins show decreasing trends again. Ganges, Brahmaputra and Indus, as well as tributaries of Lake Aral show low to medium negative trends (-1 to -20 mm yr^{-1}). The majority of Chinese and South East Asian basins show increasing trends, as do Australian basins.

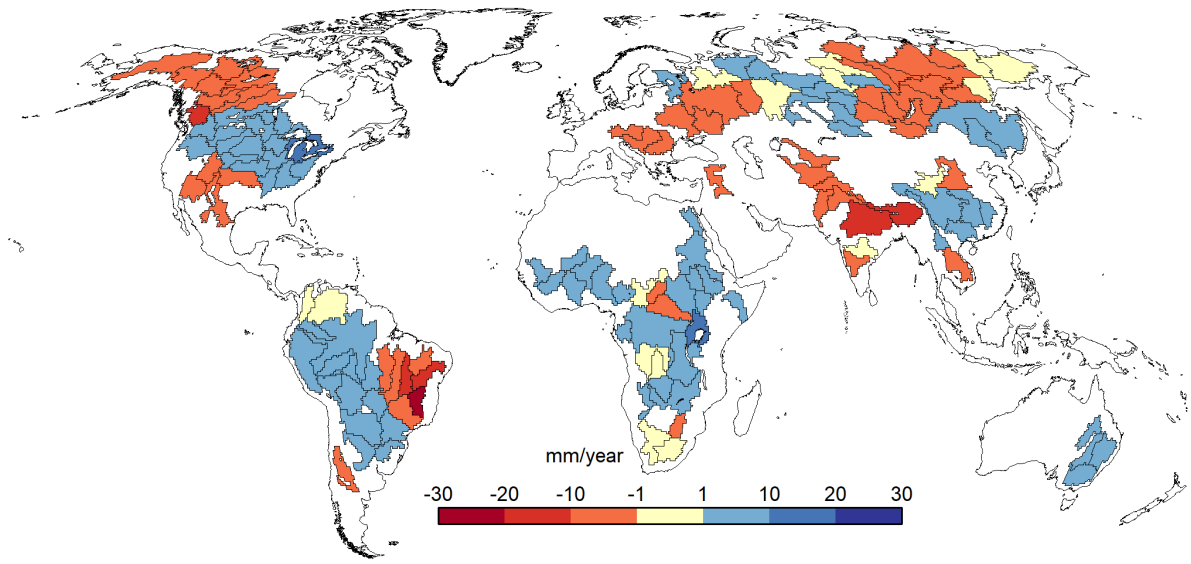

Figure 36: TWSA trend (mm yr^{-1}) in CSR-RL06 for 143 evaluated basins

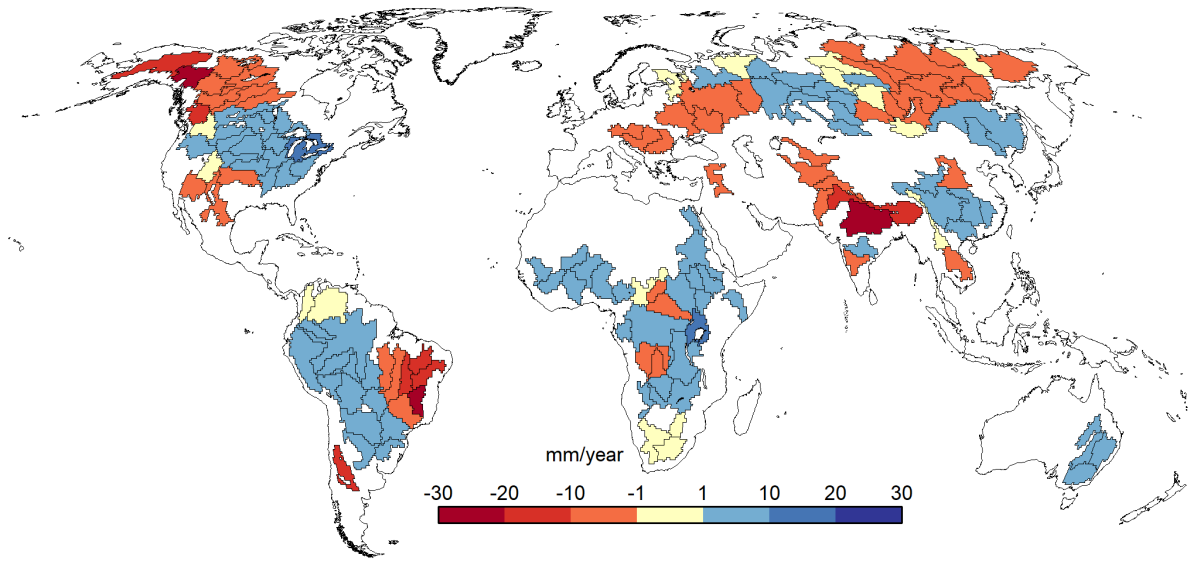

Figure 37: TWSA trend (mm yr^{-1}) in JPL-RL06M for 143 evaluated basins

Trends in model experiments

ERA5-nocal shows decreasing to stable TWSA trends for central North America (see figure 38). Decreasing trends can be identified for higher latitude basins but also for basins in the southern United States. The Amazon basin is dominated by increasing TWSA trends. However, a decreasing trend can be identified in the east of South America. The African continent is largely represented by basins showing no trend. However, the Congo basin shows low decreasing trends (-1 to -10 mm yr^{-1}). European and most Russian basins are either showing low decreasing trends or stable trends. Only far eastern Russian basins show low, increasing

trends. Basins in India and adjacent countries show largely decreasing trends, while Chinese and South East Asian basins contain positive TWSA trends.

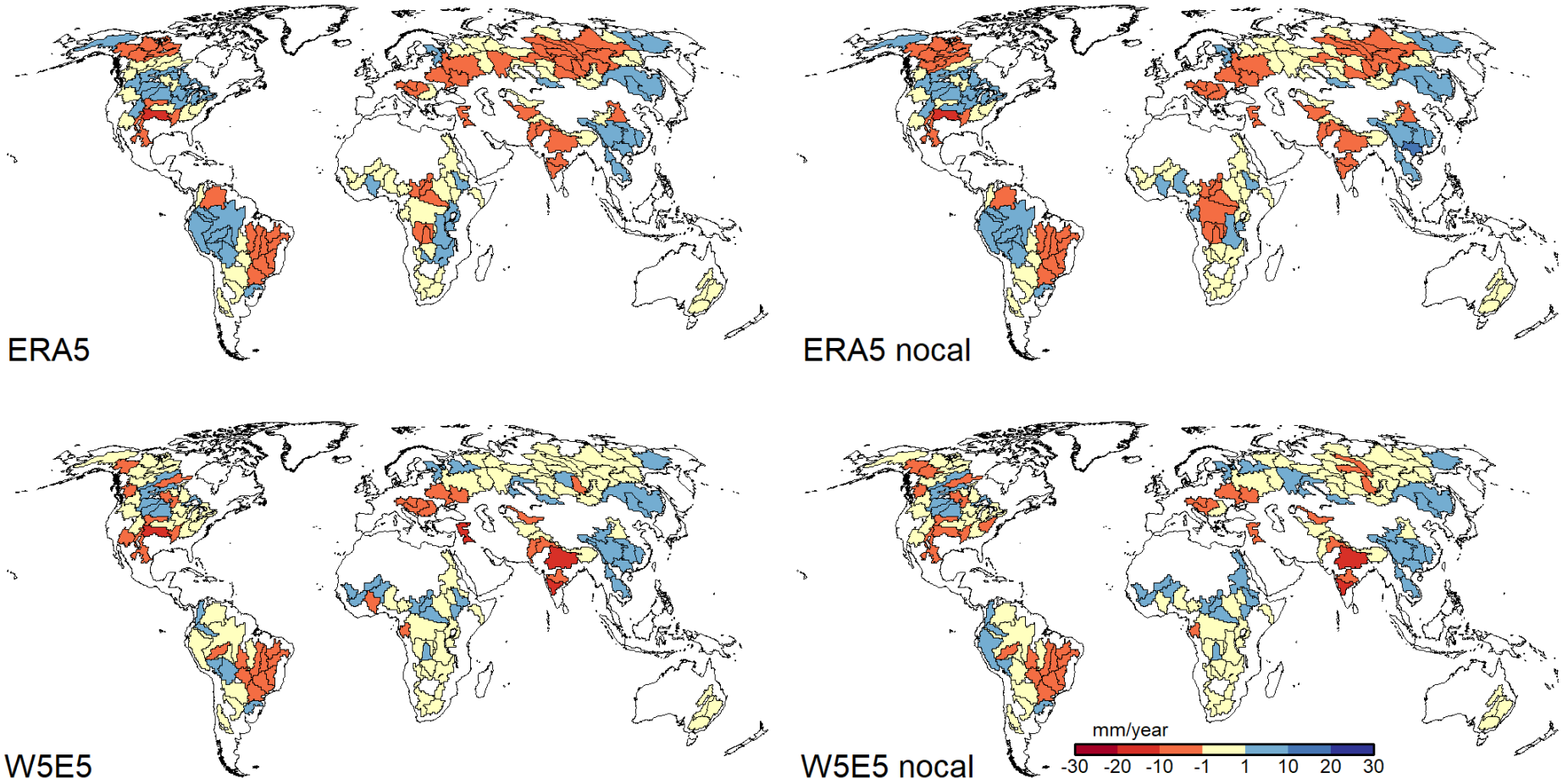

Figure 38: TWSA trend (mm yr^{-1}) of the four model experiments for 143 evaluated basins

Figure 38 shows TWSA trends in W5E5-nocal, which is dominated by basins showing no significant TWSA trend. TWSA signal in North America is rather heterogeneous, with a loose cluster of basins showing an increase of TWSA. The basins in the east of South America show a clear decreasing signal. Sub-Saharan basins and Nile tributaries tend to show low, increasing trends, while the rest of the continent shows no trend at all. However, Indian basins show low to medium decreasing trends. Increasing TWSA trends can be identified for Chinese and South East Asian basins as well.

The pattern of TWS trends remains largely the same after calibration of ERA5. However, the distribution of basins with no trend increases in North America. Additionally, the area with decreasing trends in Siberia enlarges. The centre of the Congo basin shows no trend after calibration.

Calibration does not affect general TWSA trend patterns in W5E5. The majority of evaluated basins remain to show no trend. Therefore, influences of calibration are limited to individual basins. The Amazon now includes more basins with increasing trends and more North American basins show decreasing trends.

4 Interpretation and Discussion

4.1 Update of calibration stations

As described in section 2.3 Update of Calibration Stations database, the original objective to update WaterGAP calibration stations was to find stations with equal to or preferably more than four years of discharge data after 1979. After evaluating the results of updating and extending the calibration station dataset, it became evident that the number of stations with sufficient discharge series length after 1979 was relatively small and data was often limited to the minimum requirement of four years. Shifting the core calibration period towards more recent years is beneficial since it aligns calibration and climate reanalyses time periods. However, that is not a requirement for general model success. For that reason, it had to considered whether longer calibration time series covering the distant past or shorter calibration time series covering a more recent time period are more valuable. In the case of this study, the consideration of this dilemma determined the decision against calibrating after 1979. Consequently, this meant that the success of this master thesis was again solely dependent on the re-publication and use of GSWP3. In a broader context, the search and updating process led to a general update of the WaterGAP calibration dataset, which has replaced the previously used dataset containing only 1319 stations.

In spite of the combined amount of discharge stations available from all three data sources, the number of stations incorporated into the new WaterGAP 2.2e calibration dataset was expected to be higher. During the processing of potential stations and particularly during the visual analysis in ArcMap, many GSIM stations could be identified as duplicates of GRDC stations. That itself is not a surprise since GRDC is one of the twelve databases included in GSIM and overlaps between national databases and GRDC are attributed for within the GSIM description paper. However, the authors of GSIM chose data from national databases over GRDC stations, given that national databases contained more up-to-date discharge series (DO *ET AL.*, 2018; GUDMUNDSSON *ET AL.*, 2018). Conversely, during the visual analyses of duplicate stations a different conclusion has to be drawn. For example, the broad majority of GSIM stations in Canada were removed because their GRDC duplicate had longer time series. Additionally, 203 GSIM stations in Brazil had to be removed because of implausible metadata entries such as missing station or river names. Stations that cannot be identified with high confidence are

inadequate for hydrological modelling. It is very unfortunate to exclude stations with plausible discharge data due to suspicious metadata.

The expectations regarding ADHI were comparatively high because discharge stations are unequally distributed on the African continent, which means that large areas are not represented at all. Because of ADHI's relations to SIEREM, stations are clustered in the French-speaking countries, an area fairly well covered in the previous WaterGAP calibration dataset. The resulting amount of ADHI stations is strongly influenced by the criteria formulated in the context of this work and WaterGAP calibration requirements and therefore not solely dependent on the data availability. Nevertheless, 80 ADHI stations could be used to create the new calibration dataset. 27 of these stations are located in so far unrepresented basins, which is quite an improvement.

Although the objective of enabling model calibration after 1979 was not satisfied, the outcome of the process was a success altogether. The number of calibration stations used for WaterGAP 2.2e could be increased by 190 stations improving spatial coverage by 1.3 %. In addition to an increased spatial coverage, the discharge series of already existing stations could be prolonged significantly. As spatial and temporal coverage of calibration data could be enhanced, the updating process can be judged as successful.

The calibration dataset update in addition to changes in the model structure between WaterGAP 2.2d and 2.2e limits the comparability between previous and present model results. To evaluate the influence of all or even individual changes between WaterGAP 2.2d and 2.2e is out of the scope of this master thesis. However, it would be very interesting to determine the influence and degree of improvement in a separate study.

4.2 Objective 1: Evaluation of climate forcing and calibration influences on water balance components

Precipitation of ERA5 is considerably higher than in comparable studies that evaluated water balance components (MÜLLER SCHMIED *ET AL.*, 2014, 2021; SCHNEIDER *ET AL.*, 2014). Unsurprisingly the results presented here and in CUCCHI *ET AL.* (2020) differ only marginally since the same climate forcing has been used in both evaluations. The remaining differences can be explained by the different time period that has been chosen for the study (1981-2010 in CUCCHI *ET AL.* (2020), 1979-2019 in this study). Bias adjustment with monthly GPCC precipitation totals leads to a reduction of precipitation by 7 % in W5E5 compared to ERA5.

Through bias adjustment of precipitation, the global precipitation mean of W5E5 is well within the range of precipitation values of previous studies (MÜLLER SCHMIED *ET AL.*, 2014, 2016b, 2021; CUCCHI *ET AL.*, 2020).

As a result of high precipitation in ERA5, discharge values of calibrated and uncalibrated model experiments with ERA5 are higher than those of W5E5. However, discharge in ERA5 shows insignificant differences to results presented in previous studies (MÜLLER SCHMIED *ET AL.*, 2014, 2016b, 2021). While discharge values of W5E5 are within the range of the above-mentioned studies, W5E5-nocal shows a considerably lower discharge. However, when comparing discharge results of the uncalibrated model experiments to those presented in CUCCHI *ET AL.* (2020), a striking difference, especially between discharge in ERA5, is evident even though precipitation differences are only marginal. Differences between discharge values of this study and CUCCHI *ET AL.* (2020) are attributed to the differences between the evaluated time periods with changing climatic variables and the changes between model version 2.2d to 2.2e. In order to narrow down the reasons for the remaining differences, water balance components of the time period 1981 to 2010 have been computed for 2.2e (results can be found in table 19). Nevertheless, differences in discharge remain significantly large (ERA5 difference: 1617 $km^3\ yr^{-1}$, W5E5 difference: 809 $km^3\ yr^{-1}$). Thus the dominant influence on discharge differences between CUCCHI *ET AL.* (2020) and the results presented here seem to be the result of changes in the model version. Only about 7 % of the discharge differences in ERA5 can be explained by differing climatic conditions (difference: 114 $km^3\ yr^{-1}$). Differences between discharge in W5E5 are almost entirely attributed to model changes, since less than 1 % of differences can be explained by differing climatic conditions. However, since the specific differences between CUCCHI *ET AL.* (2020) and here-presented results were only briefly evaluated, a more thorough evaluation is needed to provide sound information on individual influences.

Calibration aligns discharge values of both forcings (difference: 874 $km^3\ yr^{-1}$). The discharge alignment can be judged as an indicator for the general overestimation of discharge in ERA5 and general underestimation of discharge in W5E5. The identified tendency of both forcings in global values to either over- or underestimate discharge is supported by the spatial distribution of mean discharge evaluation of βKGE (see figure 18 and 43). After calibration, the respective tendency of either forcing is significantly reduced, just as in global discharge values. Interestingly the alignment of discharge reveals another factor. While MÜLLER SCHMIED *ET AL.* (2014) conclude that the main effect of calibration is to lower discharge, a different behaviour

can be observed here. While discharge is decreased in ERA5, it is increased in W5E5. Finally, by adjusting discharge accordingly, they are both transformed to fit into the ensemble of climate forcings producing discharges with 40000 $km^3\ yr^{-1}$and a 1000 $km^3\ yr^{-1}$ range as identified by (MÜLLER SCHMIED *ET AL.*, 2021).

Higher AET values of ERA5 compared to W5E5 result from increased precipitation and downward shortwave radiation of ERA5. Bias adjustment of precipitation in W5E5 leads to lower discharge and AET values in turn. Precipitation is the dominant driver of increased AET in ERA5-nocal and ERA5 since the precipitation differences between the climate forcings are substantial while those between PET values are marginal (0.9 %). Nevertheless, increased downward shortwave radiation positively influences ERA5s net radiation, leading to increased PET and consequently increased AET. On the other hand, while precipitation in W5E5 is already lower than in ERA5, its downward shortwave radiation is about 2.5 W/m^2 lower than that of ERA5. Lower downward shortwave radiation is the result of aerosol correction in W5E5.

As precipitation in the model is partitioned in AET and discharge, it is evident that through adjustment of the runoff coefficient and consequently discharge, AET is forced to develop into the reversed direction as discharge. Hence, a decrease in discharge accompanied by an increase in AET in ERA5 and vice versa in W5E5 can be identified. Independent of uncalibrated and calibrated model setup, all AET values presented here are considerably higher than AET values presented in other studies (MÜLLER SCHMIED *ET AL.*, 2014, 2016b). Only the AET value of W5E5 is comparable to those presented in (MÜLLER SCHMIED *ET AL.*, 2021).Table 5 in MÜLLER SCHMIED *ET AL.* (2014) lists AET results of multiple studies and enables the classification of values presented here. Without exception, all AET values of this study align with the upper end of values from the literature.

As discharge is lower here than in CUCCHI *ET AL.* (2020), AET is considerably larger. Again values from CUCCHI *ET AL.* (2020) were compared to AET computed with 2.2e for the time period 1981 to 2010. While the differences between AET results from this study and those of CUCCHI *ET AL.* (2020) amount to 1322 $km^3\ yr^{-1}$ for ERA5 and 703 $km^3\ yr^{-1}$ for W5E5, the differences between the two model versions evaluated for the same time period increases to 1629 $km^3\ yr^{-1}$ for ERA5 and 633 $km^3\ yr^{-1}$ for W5E5. Since discharge differences decreased, it is not surprising that AET differences increased. Nevertheless, the major cause for differing model results regarding AET here and in CUCCHI *ET AL.* (2020) is attributed to the changes between model versions 2.2d and 2.2e.

The highest actual water consumption (WCa) values are identified for ERA5-nocal, followed by W5E5-nocal. Both calibrated forcings show lower WCa as their uncalibrated counterpart. However, reduction of WCa in W5E5 forcing is very small (0.88 %). As high AET leads to increased irrigational water demand and WCa icludes mostly evaportation of irrigational water, it is not surprising that WCa of ERA5-nocal is high. The same is true for W5E5-nocal, which follows ERA5-nocal in the rank of AET values. In reverse, a decrease of AET consequently leads to a reduction of WCa, as seen in W5E5. Despite the identified and confirmed relationship between AET and WCa, we can see a different behavior in ERA5. ERA5 shows the highest AET values of all four evaluated model experiments but simultaneously shows the lowest WCa. Applying the above-described relationship, ERA5 should show the highest WCa, especially since AET is increased through calibration. Reduced WCa of ERA5 is mostly influenced by the reduction of actual surface water use, which can be used as a proxy for irrigational water demand (see table 3). Since crop production area is not altered between the forcings, the same water demand for crop production can be assumed. However, the source through which water demand is met may differ between the forcings. As stated above, irrigational water demand is increased with increasing AET leading to the conclusion that, if anything, irrigational water demand should be higher in both ERA5 model experiments due to higher net radiation and PET. However, since precipitation is very high in both ERA5 model experiments but actual surface water use and consequently WCa is reduced in ERA5, it is plausible that agricultural water demand is at least to a certain degree satisfied through the increased rainfalls instead of irrigation with surface and/or groundwater. Yet the influence of variations in spatial patterns of climate variables (see figure 10 - 13) should not be underestimated, but as water balance components have not been spatially disaggregated, the influence of climatic variations, especially precipitation, cannot be quantitatively analysed.

Just as AET, WCa shows increased values compared to previous evaluations of water balance compontents (MÜLLER SCHMIED *ET AL.*, 2014, 2016b, 2021). As WCa is, apart from precipitation, directly forced by temperature, variations consequently influence WCa over time. (MÜLLER SCHMIED *ET AL.*, 2016b) identified an increasing trend in global average temperature over the last three decades with 2010 as the temporal reference. Since the time period evaluated here stretches to 2019, increasing global mean temperatures have positively influenced WCa. In addition to an increase in temperature, irrigation water demand has increased steadily since 1901 with a steepening increase since the 1960 (MÜLLER SCHMIED *ET AL.*, 2016a) as a result of

increasing irrigational area, which can be seen in figure 39. Again irrigational water demand can be assumed to have grown further since 2010, thus increasing WCa here.

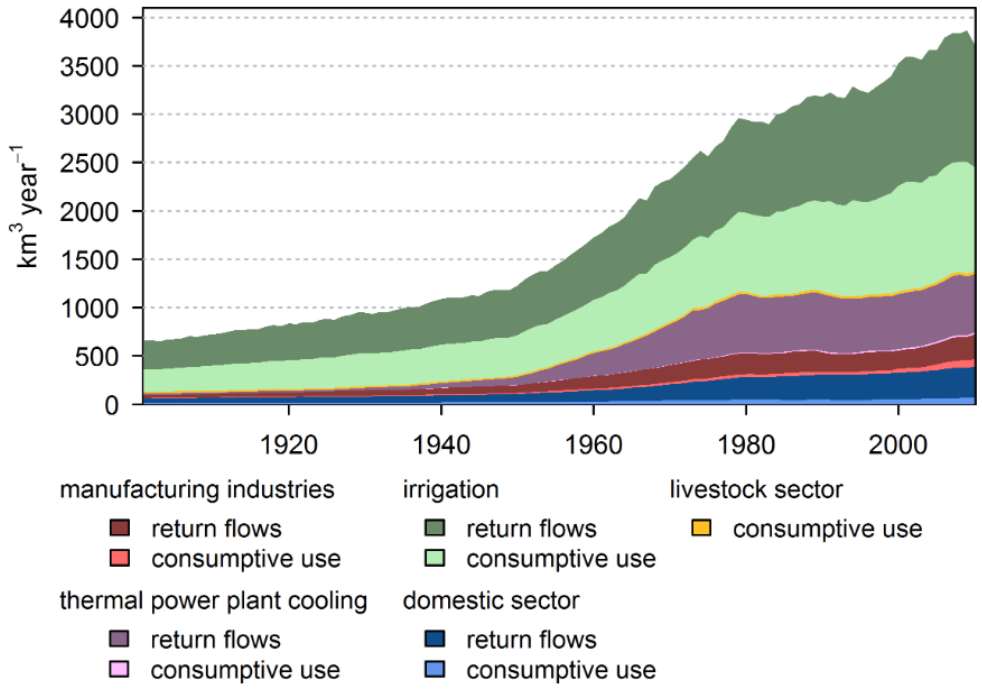

Figure 39: Development of water abstractions (sum of return flows and consumptive use) and water consumption of the five water use sectors considered in WaterGAP for 1901–2010 (MÜLLER SCHMIED *ET AL.*, 2016a)

Global change in total water storage is negative for all forcings and comparable to results in other studies (MÜLLER SCHMIED *ET AL.*, 2016b), particularly if the evaluated period is overlapping with the one evaluated here (CUCCHI *ET AL.*, 2020; MÜLLER SCHMIED *ET AL.*, 2021). Water balance error is smaller than 1 km³ yr^{-1} and therefore neglectable. The satisfying results for water balance error are comparable to those presented in MÜLLER SCHMIED *ET AL.* (2021), which means that improvements first achieved with WaterGAP version 2.2d could be sustained in version 2.2e.

4.3 Objective 2: Analyses of differences between the opimal choice of climate forcing on different spatial scales (geographic regions, climate zones and global)

In order to assess the suitability of either climate forcing, only the uncalibrated model experiments will be discussed here. Global performance is assessed by comparing the number of basins showing high performance values in the different efficiency metrics. When analysing the performance on regional to continental scales, the spatial distribution of basins with high

performance is consulted as a measure of evaluation. However, visual analyses regional to continental performances are supported by the quantative performance assessment across climate zones. Therefore, assessing model performance by the number of good performing basins, seems plausible since it is easily quantitavely analyzed. However, the size of evaluated basins varies greatly, which is why it would also be beneficial to evaluate model performance relative to the percentage of land area that reaches good performance values.

On a global scale, absolute performance of ERA5-nocal is better across almost all efficiency metrics. ERA5-nocal performs only inferior regarding the reproduction of hydrograph timing (rKGE). However, with only ten basins, differences are only marginal. Compared to results presented in CUCCHI *ET AL.* (2020), NSE values of ERA5-nocal and W5E5-nocal are not significantly different as the median is around 0. However, differences between ERA5 and WFDE5-GPCC regarding the size of the box (1. and 3. quantile) are greater in CUCCHI *ET AL.* (2020). As mentioned before, differences between results presented here and in CUCCHI *ET AL.* (2020) may stem from the chosen time period, the number of evaluation basins (1216 in CUCCHI *ET AL.* (2020), 1427 here) and their distribution.

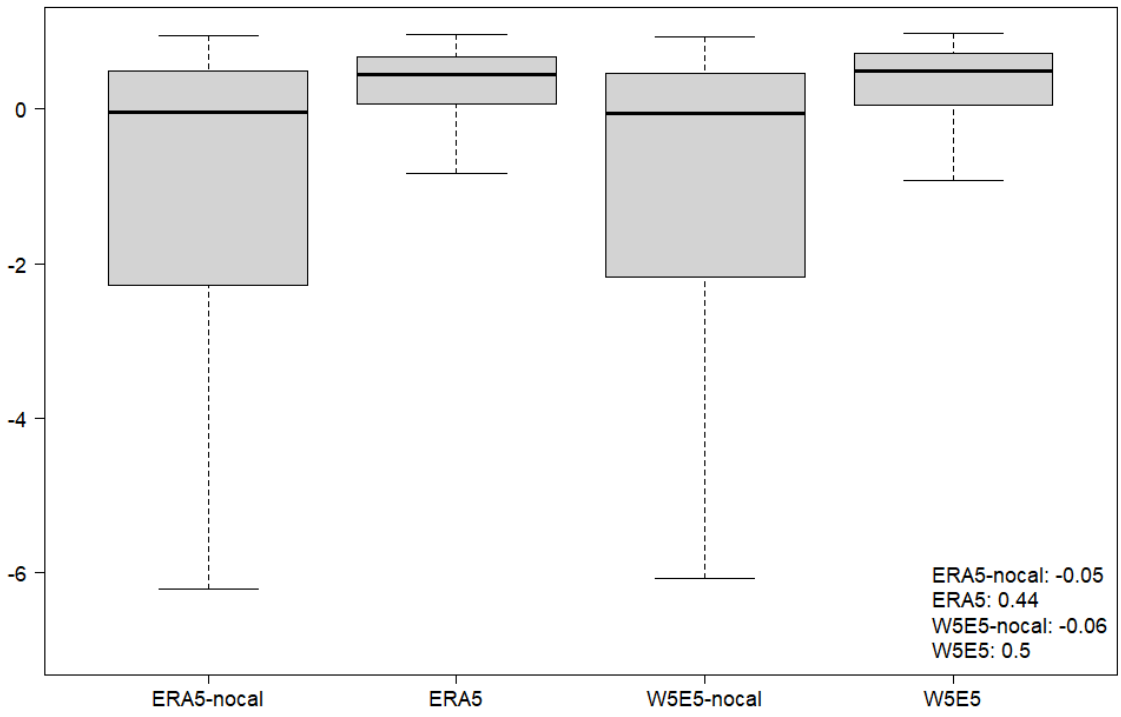

Figure 40: NSE boxplots of the four model experiments (outliers are excluded from this figure)

NSE performances with median values around 0 are not the result of poor model performance but rather of uncalibrated model runs (see figure 40). Compared to the 'nocal' model experiment in MÜLLER SCHMIED *ET AL.* (2014) results presented here are slightly better. While there are no differences in the median NSE values (equally around 0), the third quantile (lower part of the box) surpasses -4 while only extending to -2 here. The differing shapes of the third quantile box indicate a reduced spread of values included in the third quantile and simultaneously higher proximity to the median value.

The evaluation of ERA5-nocals and W5E5-nocals performance across climate zones draws a more diverse picture. While ERA5-nocal produces the best performances across all efficiency metrics in climate zone D, W5E5-nocal reaches a complementary result for climate zone A. For three (KGE, βKGE, γKGE) out of the five applied efficiency metrics, ERA5-nocal reaches higher performances in climate zone E. Performances in climate zone B are indifferent.

Although ERA5-nocal performs better in climate zone D, not all regions contribute equally to its high performance. Especially high performances in Alaska and eastern Russia lead to the overall performance superiority in climate zone D. As seen in figure 12, W5E5 shows smaller mean precipitation (difference: 100 mm yr^{-1}) in climate zone D and particular in those regions where ERA5 shows the strongest performances. The same is true for regions in climate zone E, such as the Himalayas, where ERA5-nocal tends to perform better than W5E5-nocal. Additionally, mean temperatures in climate zone D are lower in W5E5. Particularly for the Lena basin, W5E5 shows mean temperatures between -15 and -20 °C while ERA5 shows -10 to -15 °C. The combination of colder temperatures and less precipitation could lead to (1) a higher percentage of water stored as ice and (2) as a consequence of higher ice storage and less precipitation, reduced discharge. As W5E5-nocal systematically underestimates discharge in eastern Russia and Alaska, the two identified influences seem plausible.

Likewise, high performance of W5E5-nocal is not the case for all basins located in climate zone A. W5E5-nocals superiority is largely attributed to high performance in the Amazon basin. Compared to ERA5-nocal, W5E5-nocal shows reduced precipitation and a greater distribution of high temperatures (30-33 °C) over the Amazon basin. Through the bias correction with monthly precipitation means, flow dynamics of rivers in climate zone A could be better reproduced by WaterGAP.

The two climate forcings perform the worst in climate zone B. This is also true for the calibrated model experiments. Only 0 to 10 % or 1 to 11 % of all basins in climate zone B reach good

NSE or KGE values. However, climate zone B is underrepresented in both forcings. Of all 1427 evaluated basins, only 106 in ERA5 and 123 in W5E5 are located in climate zone B. Differences in climate zone definition result from differing peculiarity of climatic variables as can be seen in figure 10 - 13. The underrepresentation of basins in climate zone B is of course the result of its dry climatic characteristics which naturally restricts the occurance of streams or perennial rivers. Exceptions are rivers that originate in wetter climates such as the Nile. Nevertheless, the combination of climatic conditions reversed flow of groundwater in loosing stream regions and increased regulation cause the model to fail.

ERA5-nocal leads to better representation of streamflow across all evaluated streamflow quantiles. The spatial pattern of good streamflow representation differs, however. Satisfying streamflow representation can be frequently identified across eastern Russian basins as well as Alaskan and North American basins for the high flow indicators. Satisfying representation of low flow indicators can be found in the Amazon basin, Europe and the northern United States. Even though ERA5-nocal leads to better representation of streamflow indicators globally, W5E5-nocal cannot be classified as a poor representation for streamflow. For example, satisfying reproduction of streamflow with W5E5-nocal can be identified in the Amazon.

Linear correlation of modelled and observed TWSA does not differ between ERA5-nocal and W5E5-nocal. The only major difference in performance are the very good results for South East Asia with W5E5-nocal. Regarding TWSA, both forcings tend to underestimate trends regardless of their characteristic. Since JPL-RL06M identifies a more diverse pattern of trends, ERA5-nocal and W5E5-nocal have greater agreement with CSR-RL06. ERA5-nocal identifies more basins with trends as W5E5, which for example can be seen across northern America. However, if a trend is identified, both forcings reproduce the respective trend satisfactorily.

Finally, it can be concluded that on a global scale, ERA5 should be preferred over W5E5 as climate input, if an uncalibrated model experiement is carried out. When analyses are performed on a smaller scale like climate zones or large geographic regions (e.g. eastern Russia), the optimal choice of climate forcing depends on the chosen region. Colder dryer regions show an overall better performance with ERA5, while wetter and warmer regions are better represented with W5E5.

4.4 Objective 3: resolving whether calibration further increases the model results generated with the bias adjusted W5E5 climate forcing

Calibration increases performance of W5E5 for NSE (146 %), KGE (187 %), rKGE (6 %) and βKGE (391%). The performance of the model to capture streamflow variability (γKGE g: - 2 %) is slightly decreased through calibration. Performances of NSE and KGE overlap for most regions. Unsatisfactory results can be seen in northern America and large parts of Africa, eastern South America and central Asia. The analyses of KGE and its components allow a more thorough understanding of how these unsatisfactory performances in some regions come about. The failure to capture hydrograph timing (rKGE) is persistent for these regions in all four model experiments. Improvements through calibration are existent. However, with only 6 %, they are considerably lower than those of NSE, KGE and βKGE. In fact, calibration even leads to a small decrease in performance for basins in climate zones B and C. Why basins in northern America lead to unsatisfactory performances even though they are well observed in regard to the spatial and temporal coverage of discharge stations remains unexplained. Off timing in central Asian basins and the Nile basin, can be explained by a high degree of regulation through reservoirs and water abstractions (MÜLLER SCHMIED *ET AL.*, 2021).

Performance increases through calibration for βKGE is of no surprise since calibration fits modelled discharge to match long-term annual observed river discharge (MÜLLER SCHMIED *ET AL.*, 2021). However, the intensity of performance increases is very satisfactory and confirms the potential of W5E5. In how far the improvements of βKGE are the result of improvements between model version 2.2d and 2.2e or updating the calibration station database can not be determined. In order to analyse the individual influence of calibration database update and model improvements, further research is needed.

WaterGAP tends to underestimate variability in snow-dominated regions. A behaviour that is not exclusive of the calibrated model experiment with W5E5 but can also be identified for the remaining three model experiments. While the tendency to underestimate variability in snow-dominated basins persists after calibration, the overall performance of γKGE decreases. Performance decreases can solely be ascribed to decreases in climate zone A or the Amazon basin, to be precise. All other climate zones show performance increases. MÜLLER SCHMIED *ET AL.* (2021) identify the failure of WaterGAP to properly model wetland dynamics as the reason for unsatisfactory performances. However, since the uncalibrated set-up of WaterGAP with W5E5 leads to satisfying results, those presented in MÜLLER SCHMIED *ET AL.* (2021) are more

likely related to the choice of climate forcing and time period or even the use of a calibrated model set-up. Choice of climate forcings as it is used here summarizes changes between the different forcing versions of WFDE5 or rather W5E5, homogenization methods as well as the use of a different forcing before 1979. Since no information regarding uncalibrated model results are provided in MÜLLER SCHMIED *ET AL.* (2021), the reasons leading to the presented results cannot be narrowed down any further.

Additionally, deviations of all streamflow indicators are decreased through calibration. The number of basins with deviations up to 20 % increases by 89 % for Q1, 137 % for Q10, 94 % for Q50, 42 % for Q90 and 20 % for Q99. Improvements of Q50 can be attributed to the same reason as for βKGE. Before calibration W5E5-nocal shows insufficiencies regarding streamflow reproduction in the same regions where NSE and KGE fail. This is particularly true for the high flow indicators. However, through calibration, most streamflow indicators of the higher streamflow character show deviations below 100 %. The proportion of basins that do not surpass the mark of deviations larger than 100 % is considerably higher for low flow indicators. Again those regions with insufficient NSE and KGE values are the predominant hot-spots for basins with deviations larger than 100 %.

Most importantly, performance increases so much that W5E5 surpasses ERA5 in almost all evaluation components. The only exceptions are γKGE and complementary high and very low streamflow indicators (Q1, Q10 and Q99). Particularly strong performance increases can be seen in Asia, where basins reach the highest and very conservative category of efficiency metrics (KGE > 0.9, NSE > 0.9). Further insight of W5E5s suitability or superiority compared to ERA5 could be gained by analysing the degree of calibration that is needed to produce the here presented results. Ultimately the lower the need for more rigorous calibration steps, CFA or CFS, the higher the suitability of the climate forcing.

Unfortunately, when analysing TWSA, calibration leads to a reversed result in model performance than discharge analysis. Performance of all efficiency metrics is decreased for JPL and CSR alike. Correlation between modelled and observed TWSA is decreased through calibration. Also, for Asians basins that otherwise show extraordinary good performances. WaterGAP fails to capture the variability of TWSA in W5E5-nocal and W5E5. As described above trends, of TWSA are generally quite weak compared to those identified by JPL or CSR and calibration shows only minor effects on trend identifications (see figure 38). That calibration would reduce performance of TWSA was expected as WaterGAPs calibration

routine focuses on adjusting only one compartment of the water-cycle, namely discharge. In turn dynamics of other components are (negatively) affected as the model's performance to reproduce discharge is improved.

Even though calibration decreases WaterGAPs ability to reproduce timing and variability of TWSA, the simulation of long-term trends is only marginally affected. Long-term TWSA trends are the most interesting and relevant information that GRACE provides since they capture the development of all water storages stripped of their seasonal dynamics. The integration of long-term TWSA trends into impact modelling, could lead to better estimates of future freshwater availability and water stress. It is therefore delighting to see that calibration only marginally affects long-term trends all the while reproduction of discharge can be improved considerably.

At this point it should be noted that GRACE is prone to measurement outages, which means that the observed TWSA time series is fragmented. As TWSA time series were computed using a linear regression which omitts missing values both from the observed and the simulated TWSA time series, the resulting time series is no longer a regular time series. That is because the multiple linear regression assums equal spacing between the individual data points, which is not given anymore after stripping the time series of the months with missing data. In order to solve this problem a multiple linear regression accounting for the individual influences of linear, annual and semi-annual trends on TWSA would have been needed to approximate the monthly TWSA series of both mascon products. However, that was outside the scope of this master thesis. Furthermore, approximating the monthly TWSA series would have manipulated the data in so fare that it cannot be judged as independet anymore thus failing the objective to use GRACE TWSA data as an independent data source for model validation. Finally, it can be assumed that the general direction of trend would not change considerably. Nevertheless, precise TWSA trends computed here, should be handled with care. In order to overcome this limitation, TWSA trends have been presented in value range groups (see figure 38).

4.5 Objective 4: Assessment of W5E5s suitability for hydrological impact modelling

When analysed with an uncalibrated model-run ERA5-nocal leads to better results than W5E5-nocal, as has been analysed in detail in objective 2. Nevertheless, after calibration, performance increases of W5E5 are so strong that W5E5 surpasses ERA5, thus producing the best results out of the four model experiments. Impact Asssessments are usually performed on a regional

to basin scale, using regional hydrological models that are calibrated and validated with regard to the conditions in the study area (KRYSANOVA *ET AL.*, 2018). As calibration is an integral component of WaterGAP, allowing the model to produce the best overall results with W5E5, the claim of W5E5 to be preferably used for impact modelling is strong.

However, it is important for impact modelling to capture more than just the mean development of discharge. Thus it is necessary to evaluate the ability of a model and climate forcing to reproduce extreme flow conditions and seasonal variability as well as the compliance of hydrograph timing. Unfortunately, particularly the reproduction of extreme high and low flows (Q1, Q99) is a shortcoming of WaterGAP ran with W5E5 even after calibration. Only 39 % and 16% of all evaluated basins reach deviations limited to a maximum of 20% for Q1 and Q99 flows, respectively. Although ERA5 leads to a higher number of basins with acceptable deviations, it still fails to do so in 59 % for Q1 and 84 % for Q99 of all evaluated basins (percent difference: 2 % for Q1, < 1 % for Q99). The share of basins failing or not failing to produce satisfactory deviations for Q99 needs to be interpreted with caution due to the increased occurrence of basins with intermittent flow regimes. Where discharge drops to zero in observed data, percent deviations could not be computed due to division by zero. Differentiations between basins where division by zero occurred or deviations exceeded 100 % could not be performed. Thus the percentage of basins with satisfactory deviations can be assumed to be much higher than given here. Finally, it should be considered that in the context of this study, basins with deviations up to 20 % are classified as good performing nevertheless deviations of up to 20 % are not insignificant and can lead to large differences in discharge, particularly in the high flow regime.

Impact modelling is designed mainly to give insight into future impacts of climate and hydrological changes. Based on the provided information, policymakers can decide on adequate adaptation measures (KRYSANOVA *ET AL.*, 2018). As extreme streamflow events such as floods or the absence of discharge pose great socioeconomic risks, precise attribution of these events is at the core of assessing climate change impacts. Consequently, streamflow deviations of up to 20 % in the historical period can be classified as good when analysing performance on a global scale but are insufficient when deciding, e.g. on the scope of flood protection measures designed for the future. However that is a problem arising from the study set-up not from the quality of either climate forcings. Additionally, performances of all model experiments are of course influenced by WaterGAPs model structure and process representation (MÜLLER SCHMIED *ET AL.*, 2014, 2021; BIERKENS *ET AL.*, 2015; SCANLON *ET AL.*, 2018) (HMS 2014, 2021,

Bierkens 2015, Scanlon 2018). Therefore, W5E5s suitability for impact modelling needs to be evaluated using other models as well.

Besides W5E5's slightly inferior performance in high and low streamflow representation, it can be assumed that the overall satisfying results produced by W5E5 would be matched when used as input data for regional hydrological models. The Yangze basin seems to be the most logical choice for a first basin specific impact assessment since W5E5, combined with a calibrated model setup, leads to exceptionally good performances. The Yangtze basin is evaluated using nine subbasins, of which the majority produces good to exceptional performances across all discharge efficiency metrics. Additionally, a satisfying linear relationship of observed and modeled TWSA can be identified and TWSA variability in the Yangtze basins ranges among the best globally. TWSA trends uniformly show a 1 to 10 mm increase per year in both GRACE mascons and the W5E5 model experiment. The good reproduction of TWSA with W5E5 would even permit calibrating for TWSA.

5 Conclusion

A model-based assessment of water balance components, discharge and TWSA on different spatial scales (basin to global) was performed to evaluate the influence of (1) climate reanalyses ERA5 and its derived climate dataset W5E5, and (2) calibration on the performance of the hydrological model WaterGAP. To support the analyses and interpretation, climate variables (downward shortwave radiation, downward longwave radiation, temperature and precipitation) of both forcings were evaluated on grid cell level (figure 10- 13) and global scale (table 2). Even though the model was setup to simulate freshwater fluxes between 1901 and 2019, only the years between 1979 and 2019 were used for model performance assessments as this time period aligns with that of the climate forcing. Due to problems arising with the climate dataset GSWP3 used to prologue ERA5 as well as W5E5, WaterGAP's calibration station database was updated. The calibration dataset now comprises a total of 1509 calibration stations. Nevertheless, only 1427 stations were used to asses model performance since those stations include discharge data of at least four whole years after 1979.

The climate forcing ERA5 overestimates mean annual precipitation on a global scale, which leads to high global mean annual discharge and AET values in the respective model experiment. Bias-adjustment with monthly GPCC precipitation totals in W5E5 lead to global mean annual precipitation values that are comparable to those presented in other studies (MÜLLER SCHMIED *ET AL.*, 2014, 2016a, 2021; CUCCHI *ET AL.*, 2020). Consequently, global mean annual discharge and AET values are considerably lower in W5E5 than in ERA5. Nevertheless, when used as input to an uncalibrated model setup, ERA5 tends to overestimate while W5E5 underestimates global mean annual discharge. Discharge differences are reduced through calibration and both forcings show global man annual discharge to be 40000 $km^3\ yr^{-1}$ with a 1000 $km^3\ yr^{-1}$ range.

The ability of WaterGAP to reproduce flow dynamics was evaluated using NSE, KGE, its three components and streamflow indicators. In an uncalibrated model setup, ERA5 leads to better model performances than W5E5. Better performances of ERA5 are only marginal in regards of NSE but larger for overall KGE which is the result of strong performances in reproduction of mean discharge (βKGE) in snow-dominated regions. If the model is calibrated, W5E5 reveals its full potential by improving so much that it surpasses ERA5 in almost all efficiency metrics. The strongest performance increases with W5E5 can be seen in the models ability to reproduce

mean discharge (βKGE), particularly in regions where discharge was underestimated previously. Those regions include snow-dominated regions, which are very well reproduced with ERA5. Improvements in mean discharge reproduction of WaterGAP come at the cost of decreasing representation of variability independent of the climate forcing used as climate input. Especially, very strong performance increases in South East Asia, India and China with W5E5 are delighting to see.

The performance of WaterGAP to reproduce TWSA is limited independent of the forcing used as climate input. WaterGAP reproduces correlation and variability inadequately in the evaluated basins. The ability to identify long-term TWSA trends is only achieved for some basins. Generally, WaterGAP tends to underestimate trends in TWSA. However, for those basins where a trend is identified by the model, the signal of the trend agrees with that identified by GRACE. Calibration further reduces the performance of WaterGAP to reproduce correlation and variability. The reproduction of TWSA trends is not reduced by calibration. Ultimately, this offers the possibility to calibrate the model to long-term TWSA trends and discharge simultaneously without having to expect performance constraints on TWSA trends. The inclusion of long-term trends is a great benefit when modelling freshwater fluxes in regions with high groundwater use paired with limited monitoring capacities. Furthermore, they could be used to fit the model to most recent TWS changes (since the beginning of GRACE satellite mission in 2002) induced by climate change in order to produce more robust estimates of resource development and availability.

Finally, it can be concluded that W5E5 should be preferred as climate input for impact modelling with WaterGAP since it leads to very good model performances in a calibrated model setup. Yet depending on the spatial scale either forcing reveals certain advantages, which is why ERA5 should be at least considered as climate input. Of course, ERA5 and W5E5 can lead to different performance, when used as input for other models. Thus, it is recommended to initiate further model experiments using both climate forcings as input for different impact models. Additionally, it can be confirmed that calibration has in fact the most significant influence on WaterGAPs ability to satisfactorily model freshwater fluxes (MÜLLER SCHMIED *ET AL.*, 2014; KRYSANOVA *ET AL.*, 2018, 2020).

Bibliography

BIERKENS, M.F.P., BELL, V.A., BUREK, P., CHANEY, N., CONDON, L.E., DAVID, C.H., DE ROO, A., DÖLL, P., DROST, N., FAMIGLIETTI, J.S., FLÖRKE, M., GOCHIS, D.J., HOUSER, P., HUT, R., KEUNE, J., KOLLET, S., MAXWELL, R.M., REAGER, J.T., SAMANIEGO, L., SUDICKY, E., SUTANUDJAJA, E.H., VAN DE GIESEN, N., WINSEMIUS, H. and WOOD, E.F. (2015) 'Hyper-resolution global hydrological modelling: What is next?: "Everywhere and locally relevant"', *Hydrological Processes*, 29(2), pp. 310–320. doi:10.1002/hyp.10391.

BOERGENS, E., DOBSLAW, H., DILL, R., THOMAS, M., DAHLE, C., MURBÖCK, M. and FLECHTNER, F. (2020) 'Modelling spatial covariances for terrestrial water storage variations verified with synthetic GRACE-FO data', *GEM - International Journal on Geomathematics*, 11(1), pp. 1–25. doi:10.1007/s13137-020-00160-0.

BONAVITA, M., HÓLM, E.V., ISAKSEN, L. and FISHER, M. (2016) 'The evo-lution of the ECMWF hybrid data assimilation system', *QuarterlyJournal of the Royal Meteorological Society*, 142, pp. 287–303.

BOYER, J.F., DIEULIN, C., ROUCHÉ, N., CRES, A., SERVAT, E., PATUREL, J.-E. and MAHÉ, G. (2006) 'SIEREM an environmental information system for water resources. 5th World FRIEND Conference, La Havana - Cuba, November 2006 in Climate Variability and Change – Hydrological Impacts', *International Association of Hydrological Sciences*, 308, pp. 19–25.

C3S (2020) 'Near surface meteorological variables from 1979 to 2018 derived from bias-corrected reanalysis'. doi:10.24381/cds.20d54e34.

C3S (2021) *Climate reanalysis*. Available at: https://climate.copernicus.eu/climate-reanalysis.

CHENG, M., RIES, J.C. and TAPLEY, B.D. (2011) 'Variations of the Earth's figure axis from satellite laser ranging and GRACE', *Journal of Geophysical Research: Solid Earth*, 116(1), pp. 1–14. doi:10.1029/2010JB000850.

COMPO, G.P., WHITAKER, J.S., SARDESHMUKH, P.D., MATSUI, N., ALLAN, R.J., YIN, X., GLEASON, B.E., VOSE, R.S., RUTLEDGE, G., BESSEMOULIN, P., BRONNIMANN, S., BRUNET, M., CROUTHAMEL, R.I., GRANT, A.N., GROISMAN, P.Y., JONES, P.D., KRUK, M.C., KRUGER, A.C.,

MARSHALL, G.J., MAUGERI, M., MOK, H.Y., NORDLI, O., ROSS, T.F., TRIGO, R.M., WANG, X.L., WOODRUFF, S.D. and WORLEY, S.J. (2011) 'The Twentieth Century Reanalysis Project', *Quarterly Journal of the Royal Meteorological Society*. John Wiley and Sons Ltd, pp. 1–28. doi:10.1002/qj.776.

Copernicus Climate Change Service (2018) *ERA5 hourly data on single levels from 1979 to present - Overview*, *Copernicus*. Available at: https://www.copernicus.eu/en/access-data/copernicus-services-catalogue/era5-hourly-data-single-levels-1979-present. *Last accessed 15.12.2021*.

CUCCHI, M., WEEDON, G.P., AMICI, A., BELLOUIN, N., LANGE, S., SCHMIED, H.M., HERSBACH, H. and BUONTEMPO, C. (2020) 'WFDE5: bias adjusted ERA5 reanalysis data for impact studies Earth System Science Data Discussions'. doi:10.24381/cds.20d54e34.

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', *Quarterly Journal of the Royal Meteorological Society*, 137(656), pp. 553–597. doi:10.1002/qj.828.

DIRMEYER, P.A. (2011) 'A history and review of the Global Soil Wetness Project (GSWP)', *Journal of Hydrometeorology*, 12(5), pp. 729–749. doi:10.1175/JHM-D-10-05010.1.

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', *Bulletin of the American Meteorological Society*, 87(10), pp. 1381–1397. doi:10.1175/BAMS-87-10-1381.

DITMAR, P. (2018) 'Conversion of time-varying Stokes coefficients into mass anomalies at the Earth's surface considering the Earth's oblateness', *Journal of Geodesy*, 92(12), pp. 1401–1412. doi:10.1007/s00190-018-1128-0.

DO, H.X., GUDMUNDSSON, L., LEONARD, M. and WESTRA, S. (2018) 'The Global Streamflow Indices and Metadata Archive (GSIM)-Part 1: The production of a daily streamflow archive

and metadata', *Earth System Science Data*, 10(2), pp. 765–785. doi:10.5194/essd-10-765-2018.

DÖLL, P., HOFFMANN-DOBREV, H., PORTMANN, F.T., SIEBERT, S., EICKER, A., RODELL, M., STRASSBERG, G. and SCANLON, B.R. (2012) 'Impact of water withdrawals from groundwater and surface water on continental water storage variations', *Journal of Geodynamics*, 59–60, pp. 143–156. doi:10.1016/j.jog.2011.05.001.

DÖLL, P., KASPAR, F. and LEHNER, B. (2003) 'A global hydrological model for deriving water availability indicators: Model tuning and validation', *Journal of Hydrology*, 270(1–2), pp. 105–134. doi:10.1016/S0022-1694(02)00283-4.

DÖLL, P. and LEHNER, B. (2002) 'Validation of a new global 30-min drainage direction map', *Journal of Hydrology*, 258(1–4), pp. 214–231. doi:10.1016/S0022-1694(01)00565-0.

DÖLL, P., MÜLLER SCHMIED, H., SCHUH, C., PORTMANN, F.T. and EICKER, A. (2014) 'Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites', *Water Resources Research*, 50(7), pp. 5698–5720. doi:10.1002/2014WR015595.

DORIGO, W.A., WAGNER, W., HOHENSINN, R., HAHN, S., PAULIK, C., XAVER, A., GRUBER, A., DRUSCH, M., MECKLENBURG, S., VAN OEVELEN, P., ROBOCK, A. and JACKSON, T. (2011) 'The International Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements', *Hydrology and Earth System Sciences*, 15(5), pp. 1675–1698. doi:10.5194/hess-15-1675-2011.

EICKER, A., SCHUMACHER, M., KUSCHE, J., DÖLL, P. and SCHMIED, H.M. (2014) 'Calibration/Data Assimilation Approach for Integrating GRACE Data into the WaterGAP Global Hydrology Model (WGHM) Using an Ensemble Kalman Filter: First Results', *Surveys in Geophysics*, 35(6), pp. 1285–1309. doi:10.1007/s10712-014-9309-8.

GOSLING, S.N. and ARNELL, N.W. (2016) 'A global assessment of the impact of climate change on water scarcity', *Climatic Change*, 134(3), pp. 371–385. doi:10.1007/s10584-013-0853-x.

GRDC (2021) *Global Runoff Data Base (GRDB)*. Available at: https://www.bafg.de/GRDC/EN/Home/homepage_node.html. *Last accessed 30.12.2021*.

GRDC (no date a) ‘grdc_id_changed.txt (Download: 30.07.2021)’.see Appendix B.

GRDC (no date b) ‘GRDC station catalogue (Download: 30.07.2021)’.see Appendix B.

GUDMUNDSSON, L., DO, H.X., LEONARD, M. and WESTRA, S. (2018) ‘The Global Streamflow Indices and Metadata Archive (GSIM)-Part 2: Quality control, time-series indices and homogeneity assessment’, *Earth System Science Data*, 10(2), pp. 787–804. doi:10.5194/essd-10-787-2018.

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’, *Journal of Hydrology*, 377(1–2), pp. 80–91. doi:10.1016/j.jhydrol.2009.08.003.

GUPTA, H.V., SOROOSHIAN, S. and YAPO, P.O. (1998) ‘Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information’, *Water Resources Research*, 34(4), pp. 751–763. doi:10.1029/97WR03495.

HARRIS, I., JONES, P.D., OSBORN, T.J. and LISTER, D.H. (2014) ‘Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset’, *International Journal of Climatology*, 34(3), pp. 623–642. doi:10.1002/joc.3711.

HERSBACH, H., BELL, B., BERRISFORD, P., HIRAHARA, S., HORÁNYI, A., MUÑOZ-SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R., SCHEPERS, D., SIMMONS, A., SOCI, C., ABDALLA, S., ABELLAN, X., BALSAMO, G., BECHTOLD, P., BIAVATI, G., BIDLOT, J., BONAVITA, M., DE CHIARA, G., DAHLGREN, P., DEE, D., DIAMANTAKIS, M., DRAGANI, R., FLEMMING, J., FORBES, R., FUENTES, M., GEER, A., HAIMBERGER, L., HEALY, S., HOGAN, R.J., HÓLM, E., JANISKOVÁ, M., KEELEY, S., LALOYAUX, P., LOPEZ, P., LUPU, C., RADNOTI, G., DE ROSNAY, P., ROZUM, I., VAMBORG, F., VILLAUME, S. and THÉPAUT, J.N. (2020) ‘The ERA5 global reanalysis’, *Quarterly Journal of the Royal Meteorological Society*, 146(730), pp. 1999–2049. doi:10.1002/qj.3803.

HIRABAYASHI, Y., KANAE, S., MOTOYA, K., MASUDA, K. and DÖLL, P. (2008) ‘A 59-year (1948-2006) global meteorological forcing data set for land surface models. Part II: Global snowfall estimation’, *Hydrological Research Letters*, 2, pp. 65–69. doi:10.3178/hrl.2.65.

HUNGER, M. and DÖLL, P. (2008) *Value of river discharge data for global-scale hydrological modeling*, *Hydrol. Earth Syst. Sci.* Available at: www.hydrol-earth-syst-sci.net/12/841/2008/.

ISIMIP (2021a) *ISIMIP3 Protocol*. Available at: https://www.isimip.org/protocol/3/. *Last accessed 28.11.2021.*

ISIMIP (2021b) 'ISIMIP3a: climate input data update', 13 September. Available at: https://www.isimip.org/gettingstarted/input-data-changelog/isimip3a-climate-input-data-update/. *Last accessed 28.11.2021.*

ISIMIP (2021c) 'ISIMIP3a: Discontinuities found in climate input data', 9 March. Available at: https://www.isimip.org/gettingstarted/input-data-changelog/isimip3a-discontinuities-found-climate-input-data/. *Last accessed 28.11.2021.*

Jet Propulsion Laboratory (2021a) *GRACE & GRACE-FO - Data Months / Days*. Available at: https://grace.jpl.nasa.gov/data/grace-months/. *Last accessed 14.11.2021.*

Jet Propulsion Laboratory (2021b) *GRACE Tellus*. Available at: https://grace.jpl.nasa.gov/. *Last accessed 14.11.2021.*

KAUFFELDT, A., HALLDIN, S., RODHE, A., XU, C.Y. and WESTERBERG, I.K. (2013) 'Disinformative data in large-scale hydrological modelling', *Hydrology and Earth System Sciences*, 17(7), pp. 2845–2857. doi:10.5194/hess-17-2845-2013.

KIM, H. (2014) *Global Soil Wetness Project Phase 3*. Available at: http://hydro.iis.u-tokyo.ac.jp/GSWP3/exp1.html. *Last accessed 28.11.2021.*

KIM, H. (2017) 'Global Soil Wetness Project Phase 3 Atmospheric Bound-ary Conditions (Experiment 1), Data Integration and AnalysisSystem [data set]'. doi:https://doi.org/10.20783/DIAS.501.

KLING, H., FUCHS, M. and PAULIN, M. (2012) 'Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios', *Journal of Hydrology*, 424–425, pp. 264–277. doi:10.1016/j.jhydrol.2012.01.011.

KRAUSE, P., BOYLE, D.P. and BÄSE, F. (2005) 'Comparison of different efficiency criteria for hydrological model assessment', *Advances in Geosciences*, 5, pp. 89–97. doi:10.5194/adgeo-5-89-2005.

KRYSANOVA, V., DONNELLY, C., GELFAN, A., GERTEN, D., ARHEIMER, B., HATTERMANN, F. and KUNDZEWICZ, Z.W. (2018) 'How the performance of hydrological models relates to

credibility of projections under climate change', *Hydrological Sciences Journal*, 63(5), pp. 696–720. doi:10.1080/02626667.2018.1446214.

KRYSANOVA, V., ZAHERPOUR, J., DIDOVETS, I., GOSLING, S.N., GERTEN, D., HANASAKI, N., MÜLLER SCHMIED, H., POKHREL, Y., SATOH, Y., TANG, Q. and WADA, Y. (2020) 'How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change', *Climatic Change*, 163(3), pp. 1353–1377. doi:10.1007/s10584-020-02840-0.

LANGE, S. (2019) 'Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0)', *Geoscientific Model Development*, 12(7). doi:10.5194/gmd-12-3055-2019.

LANGE, S. (2021) 'ISIMIP3BASD v2.5.0 [data set]'. doi:10.5281/zenodo.4686991.

LANGE, S., MENZ, C., GLEIXNER, S., CUCCHI, M., WEEDON, G.P., AMICI, A., BELLOUIN, N., MÜLLER SCHMIED, H., HERSBACH, H., BUONTEMPO, C. and CAGNAZZO, C. (2021) 'WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0)'. doi:10.48364/ISIMIP.342217.

LEGATES, D.R. and MCCABE, G.J. (1999) 'Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation', *Water Resources Research*, 35(1), pp. 233–241. doi:10.1029/1998WR900018.

LOOMIS, B.D., RACHLIN, K.E. and LUTHCKE, S.B. (2019) 'Improved Earth Oblateness Rate Reveals Increased Ice Sheet Losses and Mass-Driven Sea Level Rise', *Geophysical Research Letters*, 46(12), pp. 6910–6917. doi:10.1029/2019GL082929.

MENGEL, M., TREU, S., LANGE, S. and FRIELER, K. (2021) 'ATTRICI 1.0 – counterfactual climate for impact attribution', *Geoscientific Model Development Discussions*, 14, pp. 5269–5284. doi:10.5194/gmd-2020-145.

MITCHELL, T.D. and JONES, P.D. (2005) 'An improved method of constructing a database of monthly climate observations and associated high-resolution grids', *International Journal of Climatology*, 25(6), pp. 693–712. doi:10.1002/joc.1181.

MORIASI, D.N., GITAU, M.W., PAI, N. and DAGGUPATI, P. (2015) 'Hydrologic and water quality models: Performance measures and evaluation criteria', *Transactions of the ASABE*, 58(6), pp. 1763–1785. doi:10.13031/trans.58.10715.

MÜLLER SCHMIED, H., ADAM, L., EISNER, S., FINK, G., FLÖRKE, M., KIM, H., OKI, T., PORTMANN, F.T., REINECKE, R., RIEDEL, C., SONG, Q., ZHANG, J. and DÖLL, P. (2016a) 'Impact of climate forcing uncertainty and human water use on global and continental water balance components', in *Proceedings of the International Association of Hydrological Sciences*. Copernicus GmbH, pp. 53–62. doi:10.5194/piahs-374-53-2016.

MÜLLER SCHMIED, H., ADAM, L., EISNER, S., FINK, G., FLÖRKE, M., KIM, H., OKI, T., PORTMANN, F.T., REINECKE, R., RIEDEL, C., SONG, Q., ZHANG, J. and DÖLL, P. (2016b) 'Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use', *Hydrology and Earth System Sciences*, 20(7), pp. 2877–2898. doi:10.5194/hess-20-2877-2016.

MÜLLER SCHMIED, H., CACERES, D., EISNER, S., FLÖRKE, M., HERBERT, C., NIEMANN, C., ASALI PEIRIS, T., POPAT, E., THEODOR PORTMANN, F., REINECKE, R., SCHUMACHER, M., SHADKAM, S., TELTEU, C.E., TRAUTMANN, T. and DÖLL, P. (2021) 'The global water resources and use model WaterGAP v2.2d: Model description and evaluation', *Geoscientific Model Development*, 14(2), pp. 1037–1079. doi:10.5194/gmd-14-1037-2021.

MÜLLER SCHMIED, H., EISNER, S., FRANZ, D., WATTENBACH, M., PORTMANN, F.T., FLÖRKE, M. and DÖLL, P. (2014) 'Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration', *Hydrology and Earth System Sciences*, 18(9), pp. 3511–3538. doi:10.5194/hess-18-3511-2014.

NASH, J.E. and SUTCLIFFE, J. V. (1970) 'No 1970 River flow forecasting throughconceptual models part I—A discussion of principles', *Journal of Hydrology*, 10, pp. 282–290.

REINECKE, R., MÜLLER SCHMIED, H., TRAUTMANN, T., SEABY ANDERSEN, L., BUREK, P., FLÖRKE, M., GOSLING, S.N., GRILLAKIS, M., HANASAKI, N., KOUTROULIS, A., POKHREL, Y., THIERY, W., WADA, Y., YUSUKE, S. and DÖLL, P. (2021) 'Uncertainty of simulated groundwater recharge at different global warming levels: A global-scale multi-model ensemble study', *Hydrology and Earth System Sciences*, 25(2), pp. 787–810. doi:10.5194/hess-25-787-2021.

RICHARD PELTIER, W., ARGUS, D.F. and DRUMMOND, R. (2018) 'Comment on "An Assessment of the ICE-6G_C (VM5a) Glacial Isostatic Adjustment Model" by Purcell et al.', *Journal of Geophysical Research: Solid Earth*, 123(2), pp. 2019–2028. doi:10.1002/2016JB013844.

SAVE, H. (2020) ‘CSR GRACE and GRACE-FO RL06 Mascon Solutions v02’. doi:10.15781/cgq9-nh24.

SAVE, H., BETTADPUR, S. and TAPLEY, B.D. (2016) ‘High-resolution CSR GRACE RL05 mascons’, *Journal of Geophysical Research: Solid Earth*, 121(10), pp. 7547–7569. doi:10.1002/2016JB013007.

SCANLON, B.R., ZHANG, Z., SAVE, H., SUN, A.Y., SCHMIED, H.M., VAN BEEK, L.P.H., WIESE, D.N., WADA, Y., LONG, D., REEDY, R.C., LONGUEVERGNE, L., DÖLL, P. and BIERKENS, M.F.P. (2018) ‘Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(6), pp. E1080–E1089. doi:10.1073/pnas.1704665115.

SCHEWE, J., HEINKE, J., GERTEN, D., HADDELAND, I., ARNELL, N.W., CLARK, D.B., DANKERS, R., EISNER, S., FEKETE, B.M., COLÓN-GONZÁLEZ, F.J., GOSLING, S.N., KIM, H., LIU, X., MASAKI, Y., PORTMANN, F.T., SATOH, Y., STACKE, T., TANG, Q., WADA, Y., WISSER, D., ALBRECHT, T., FRIELER, K., PIONTEK, F., WARSZAWSKI, L. and KABAT, P. (2014) ‘Multimodel assessment of water scarcity under climate change’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(9), pp. 3245–3250. doi:10.1073/pnas.1222460110.

SCHMIDT, R., SCHWINTZER, P., FLECHTNER, F., REIGBER, C., GÜNTNER, A., DÖLL, P., RAMILLIEN, G., CAZENAVE, A., PETROVIC, S., JOCHMANN, H. and WÜNSCH, J. (2006) ‘GRACE observations of changes in continental water storage’, *Global and Planetary Change*, 50(1–2), pp. 112–126. doi:10.1016/j.gloplacha.2004.11.018.

SCHNEIDER, U., BECKER, A., FINGER, P., MEYER-CHRISTOFFER, A. and ZIESE, M. (2011) ‘GPCC Full Data Monthly Product Ver-sion 2018 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historical Data’. doi:https://doi.org/10.5676/DWD_GPCC/FD_M_V2018_050.

SCHNEIDER, U., BECKER, A., FINGER, P., MEYER-CHRISTOFFER, A., ZIESE, M. and RUDOLF, B. (2014) ‘GPCC’s new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle’, *Theoretical and Applied Climatology*, 115(1–2), pp. 15–40. doi:10.1007/s00704-013-0860-x.

SCHUMACHER, M., FOROOTAN, E., VAN DIJK, A.I.J.M., SCHMIED, H.M., CROSBIE, R.S.,

KUSCHE, J. and DÖLL, P. (2018) ‘Improving drought simulations within the Murray-Darling Basin by combined calibration/assimilation of GRACE data into the WaterGAP Global Hydrology Model’, *Remote Sensing of Environment*, 204, pp. 212–228.

SHEFFIELD, J., WOOD, E.F., PAN, M., BECK, H., COCCIA, G., SERRAT-CAPDEVILA, A. and VERBIST, K. (2018) ‘Satellite Remote Sensing for Water Resources Management: Potential for Supporting Sustainable Development in Data-Poor Regions’, *Water Resources Research*, 54(12), pp. 9724–9758. doi:10.1029/2017WR022437.

STACKHOUSE, P.W., GUPTA, S.K., COX, S.J., ZHANG, T., MIKOVITZ, J.C. and HINKELMAN, L.M. (2011) ‘24.5-year SRB data set released’, *GEWEX News*, 21(1), pp. 10–12.

SUN, Y., RIVA, R. and DITMAR, P. (2016) ‘Optimizing estimates of annual variations and trends in geocenter motion and J2 from a combination of GRACE data and geophysical models’, *Journal of Geophysical Research: Solid Earth*, 121(11), pp. 8352–8370. doi:10.1002/2016JB013073.

SWENSON, S., CHAMBERS, D. and WAHR, J. (2008) ‘Estimating geocenter variations from a combination of GRACE and ocean model output’, *Journal of Geophysical Research: Solid Earth*, 113(8), pp. 1–12. doi:10.1029/2007JB005338.

TANGDAMRONGSUB, N., STEELE-DUNNE, S.C., GUNTER, B.C., DITMAR, P.G. and WEERTS, A.H. (2015) ‘Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin’, *Hydrology and Earth System Sciences*, 19(4), pp. 2079–2100. doi:10.5194/hess-19-2079-2015.

TRAMBLAY, Y., ROUCHÉ, N., PATUREL, J.E., MAHÉ, G., BOYER, J.F., AMOUSSOU, E., BODIAN, A., DACOSTA, H., DAKHLAOUI, H., DEZETTER, A., HUGHES, D., HANICH, L., PEUGEOT, C., TSHIMANGA, R. and LACHASSAGNE, P. (2021) ‘ADHI: The African Database of Hydrometric Indices (1950-2018)’, *Earth System Science Data*, 13(4), pp. 1547–1560. doi:10.5194/essd-13-1547-2021.

United Nations (2018) *Sustainable Development Goal 6 Synthesis Report on Water and Sanitation, United Nations*.

VELDKAMP, T.I.E., ZHAO, F., WARD, P.J., DE MOEL, H., AERTS, J.C.J.H., SCHMIED, H.M., PORTMANN, F.T., MASAKI, Y., POKHREL, Y., LIU, X., SATOH, Y., GERTEN, D., GOSLING, S.N.,

ZAHERPOUR, J. and WADA, Y. (2018) 'Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: A multi-model validation study', *Environmental Research Letters*. Institute of Physics Publishing. doi:10.1088/1748-9326/aab96f.

WADA, Y., VAN BEEK, L.P.H. and BIERKENS, M.F.P. (2012) 'Nonsustainable groundwater sustaining irrigation: A global assessment', *Water Resources Research*, 48(1). doi:10.1029/2011WR010562.

WADA, Y., VAN BEEK, L.P.H., VAN KEMPEN, C.M., RECKMAN, J.W.T.M., VASAK, S. and BIERKENS, M.F.P. (2010) 'Global depletion of groundwater resources', *Geophysical Research Letters*, 37(20), pp. 1–5. doi:10.1029/2010GL044571.

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', *Water Resources Research*, 50(9), pp. 7505–7514. doi:10.1002/2014WR015638.

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', *Journal of Hydrometeorology*, 12(5), pp. 823–848. doi:10.1175/2011JHM1369.1.

WIESE, D.N., LANDERER, F.W. and WATKINS, M.M. (2016) 'Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution', *Water Resources Research*, (52), pp. 7490–7502. doi:10.1002/2016WR019344.Received.

WIESE, D.N., YUAN, D.-N., BOENING, C., LANDERER, F.W. and WATKINS, M.M. (2019) 'JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL06 CRI Filtered Version 02. Ver. 02. PO.DAAC'. doi:10.5067/TEMSC-3JC62.

YOSHIMURA, K., KANAMITSU, M., NOONE, D. and OKI, T. (2008) 'Historical isotope simulation using Reanalysis atmospheric data', *Journal of Geophysical Research Atmospheres*, 113(19), pp. 1–15. doi:10.1029/2008JD010074.

ZAMBRANO-BIGIARINI, M. (2020) 'Package "hydroGOF"', pp. 1–77. doi:10.1002/hyp.7072.

Appendix A

A.1 Additional Water Balance Components

Table 19: Global water balance components (excluding Antarctica and Greenland) for 1981 to 2010. All units in $km^3\ yr^{-1}$. Actual evapotranspiration includes actual consumptive water use. Actual consumptive use is the sum of row 5 and 6. Long-term average volume balance error is computed as the difference of precipitation and the sum of components 2, 4 and 8

No.	Component	ERA5-nocal	ERA5	W5E5-nocal	W5E5
1	Precipitation	120244	120244	111351	111351
2	Streamflow into oceans and inland sinks	42006	40462	37326	39568
3	Potential evapotranspiration	149867	149867	148704	148707
4	Actual evapotranspiration	78324	79886	74063	71841
5	Actual net abstraction from surface water	1640	1473	1533	1518
6	Actual net abstraction from groundwater	-98	-86	-100	-94
7	Actual consumptive water use	1542	1387	1432	1424
8	Change of total water storage	-86	-104	-38	-58
9	Long-term average volume balance error	-0.25	-0.24	-0.23	-0.21

A.2 Additional Efficiency metrics

A.2.1 Boxplots of Efficiency metrics

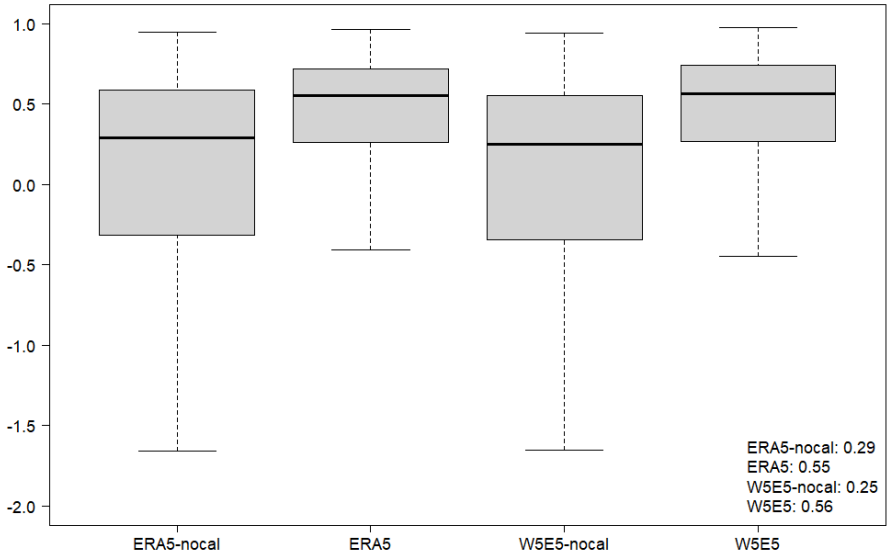

Figure 41: KGE boxplots of the four model experiments

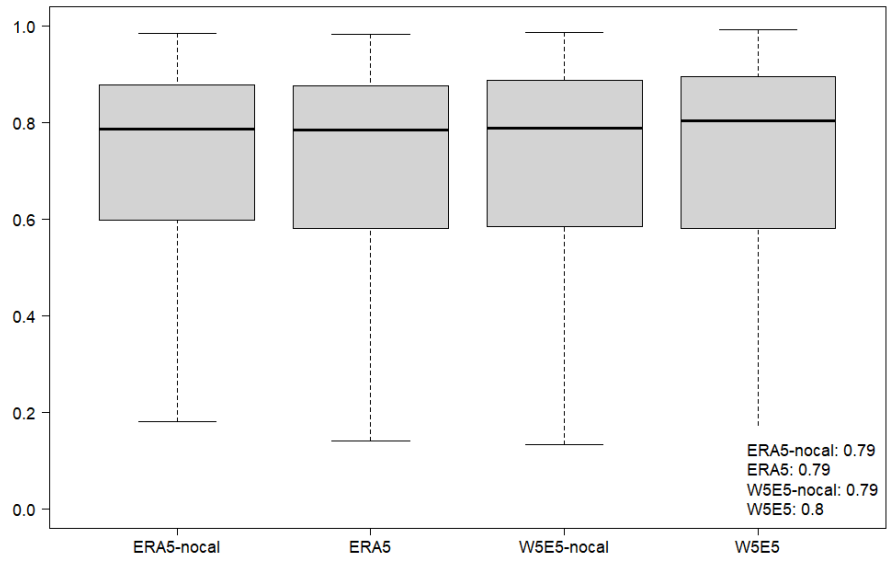

Figure 42: rKGE boxplots of the four model experiments

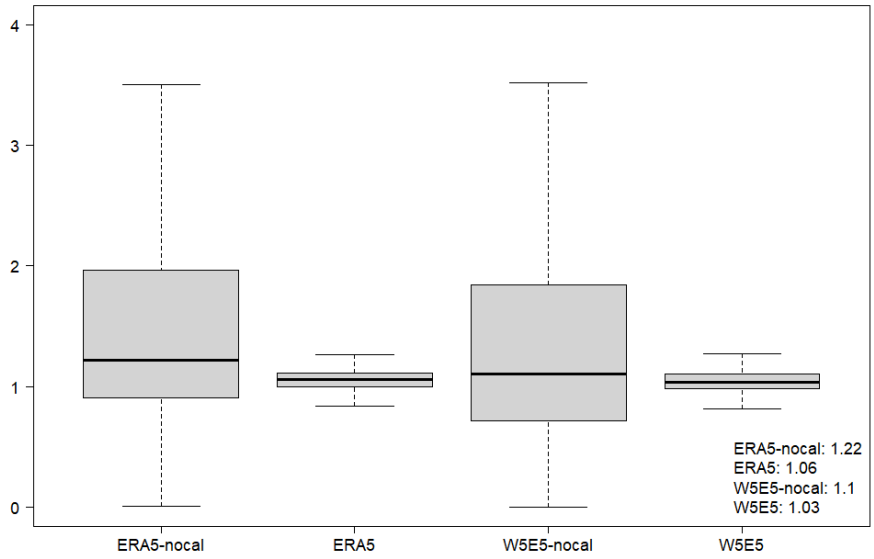

Figure 43: βKGE boxplots of the four model experiments

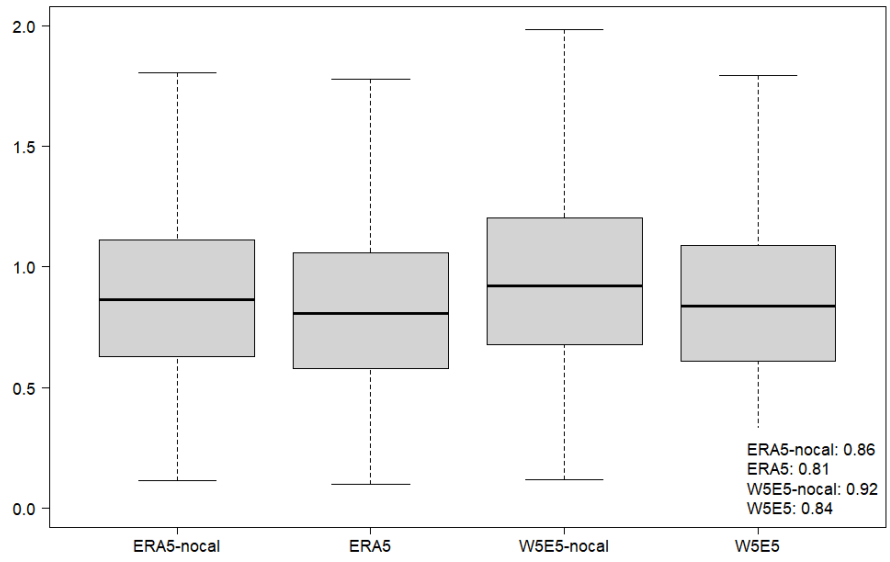

Figure 44: γKGE boxplots of the four model experiments

A.2.2 Percent bias

Percent bias (PBIAS) evaluates the average tendency of simulated data to deviate from observed values offering the opportunity to differentiate between the tendency of simulated values to be larger or smaller (GUPTA *ET AL.*, 1998). The aspired value for PBIAS is zero, however values only slightly deviating from zero already indicate high accuracy of a model. A model tends to have smaller values if PBIAS is positive and vice versa. If a model overpredicts as much as it underpredicts, the resulting PBIAS will be zero indicating a good model

performance. Just like other performance indicators PBIAS should therefore be used with other statistical measures (MORIASI *ET AL.*, 2015). PBIAS expressed as percentage is calculated as:

$$PBIAS = \left[\frac{\sum_{i=1}^{n}(O_i - S_i) * 100}{\sum_{i=1}^{n}(O_i)}\right] \quad (8)$$

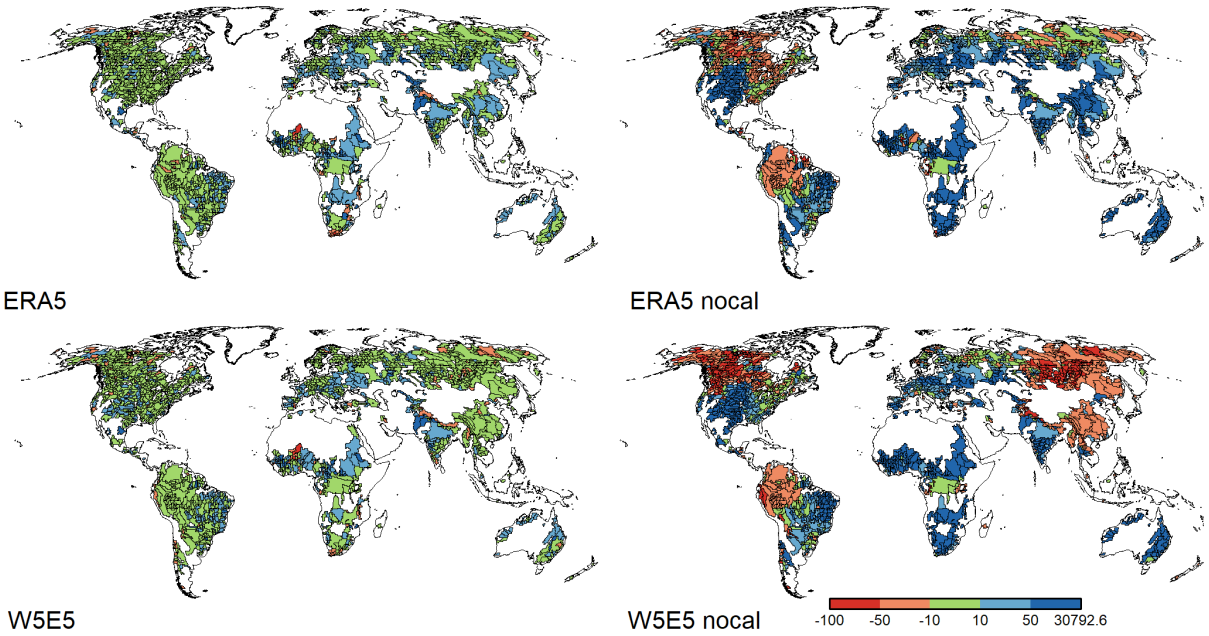

Figure 45: PBIAS of the four model experiments evaluated for 1427 basins

A.3 Additional Streamflow Indicator

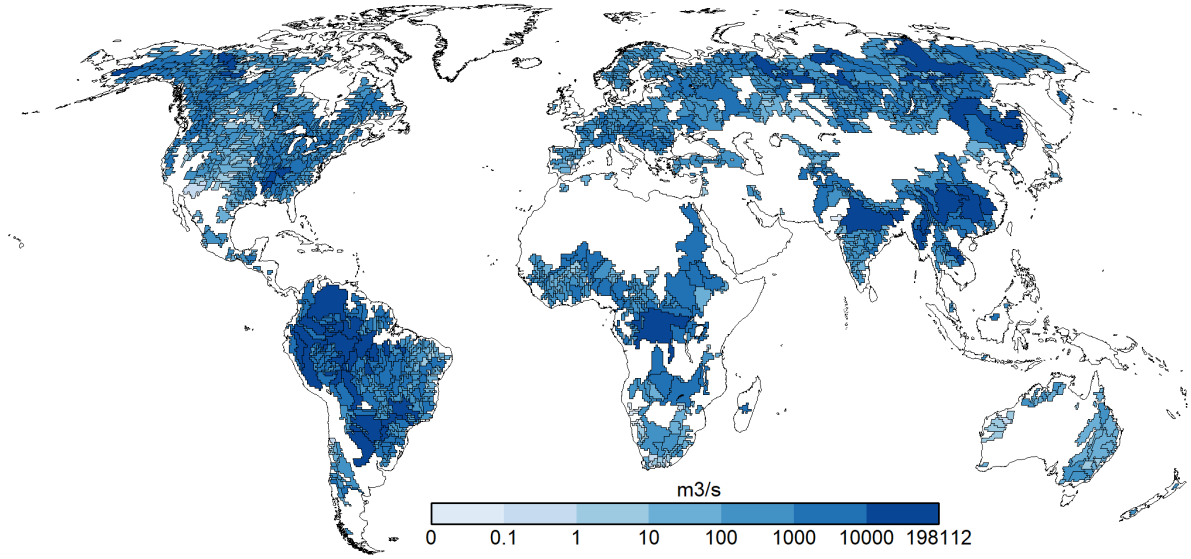

Figure 46: Q25 streamflows ($m^3\ s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

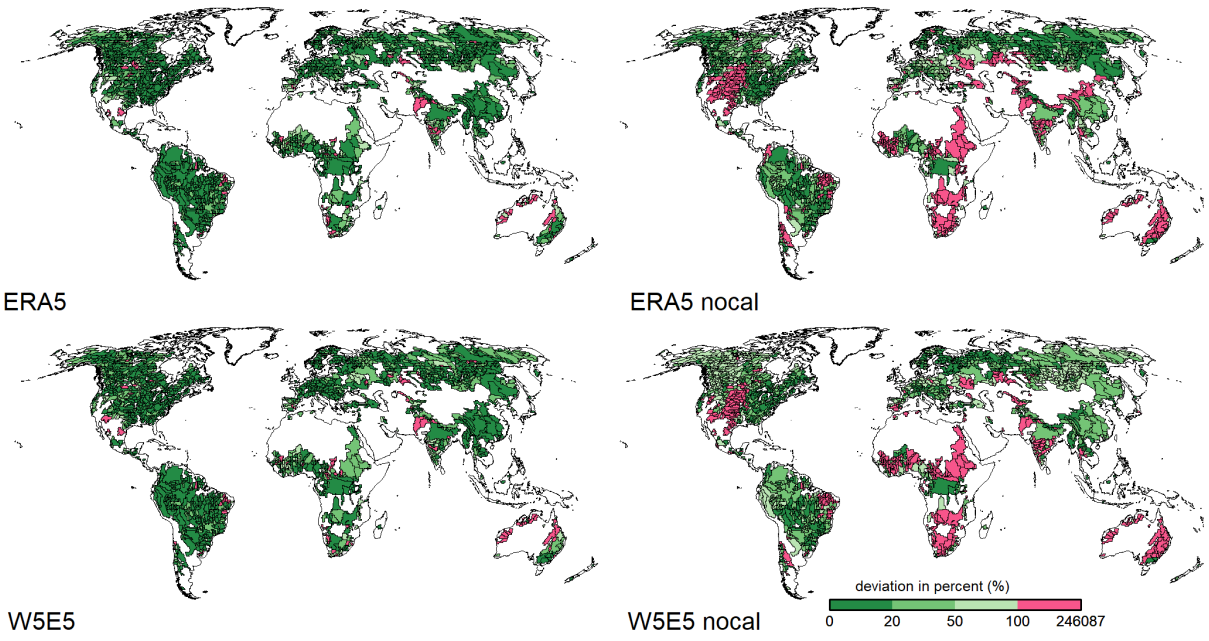

Figure 47: Deviations (%) of modelled O25 flows from observed Q25 flows for 1427 basins

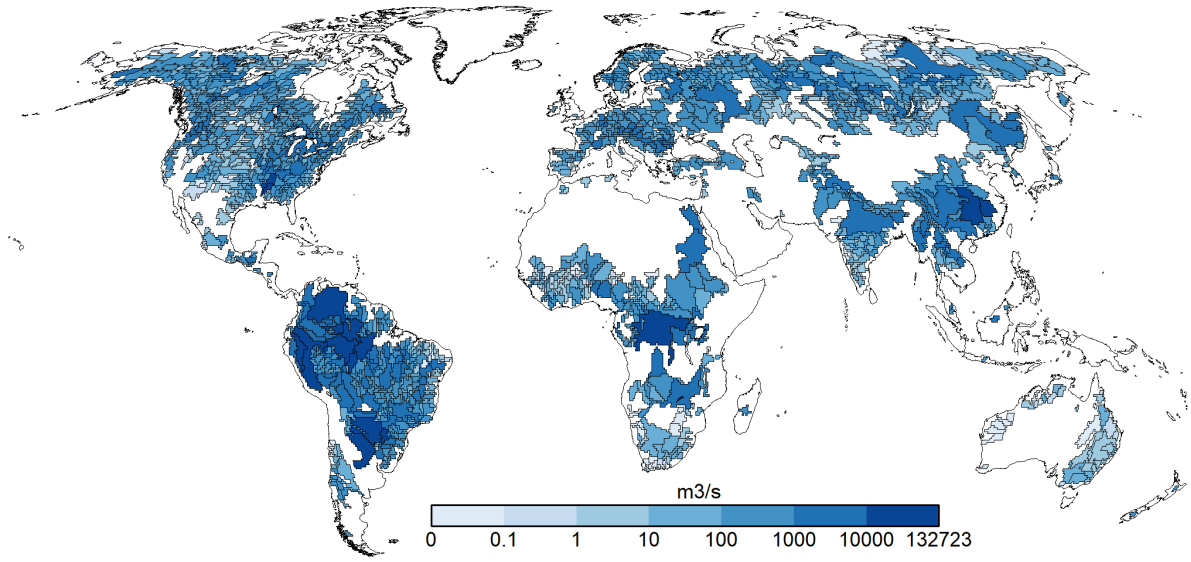

Figure 48: Q75 streamflows ($m^3 s^{-1}$) at 1427 stations evaluated for the period 1979 to 2019

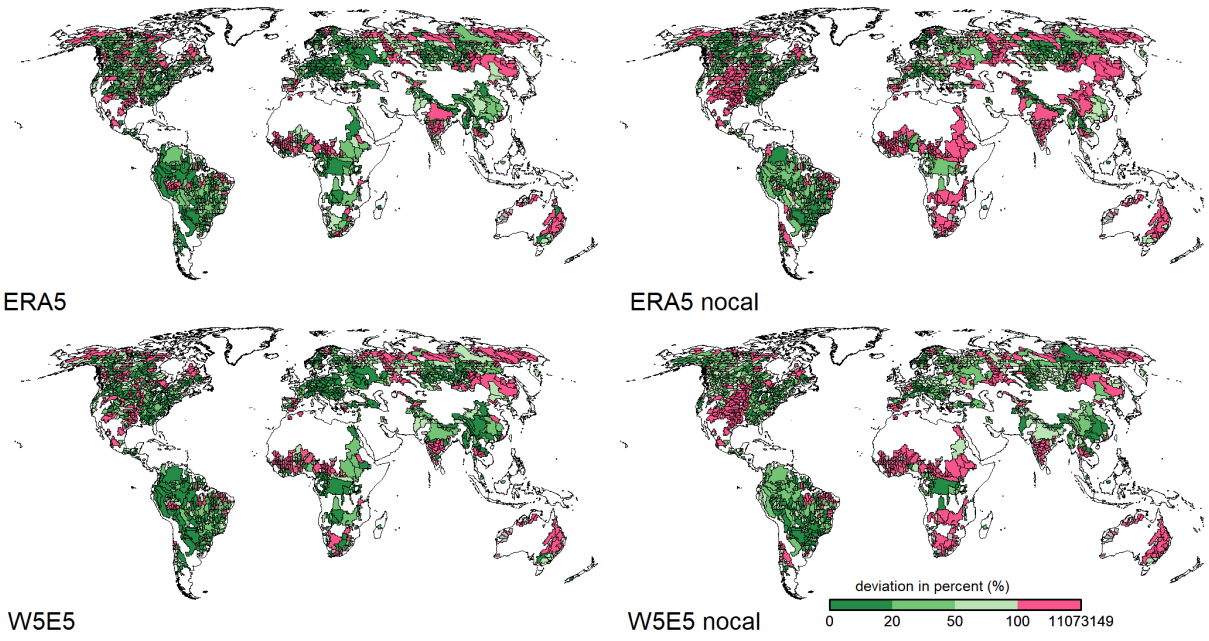

Figure 49: Deviations (%) of modelled O75 flows from observed Q75 flows for 1427 basins

Appendix B

B.1 Calibration Station Update

Subfolder ADHI

This folder contains all discharge data provided by ADHI, the scripts to evaluate ADHI discharge data as well as the final set of discharge data used for WaterGAP 2.2e calibration dataset. Furthermore, the ADHI description paper can be found within the folder (TRAMBLAY *ET AL.*, 2021).

Subfolder GRDC

This folder contains all discharge data provided by GRDC, the scripts to identify the updated station IDs, the scripts evaluate GRDC discharge data as well as the final set of discharge data used for WaterGAP 2.2e calibration dataset.

Subfolder GSIM

This folder contains all discharge data provided by GSIM, the scripts to evaluate GSIM discharge data as well as the final set of discharge data used for WaterGAP 2.2e calibration

dataset. Furthermore, the GSIM description paper can be found within the folder (DO *ET AL.*, 2018; GUDMUNDSSON *ET AL.*, 2018).

B.2 Model Experiments

Subfolder climate

This folder includes all evaluations of climate variables (downward shortwave and longwave radiation, precipitation and temperature) as well as the derivation of climate zones for ERA5 and W5E5.

Subfolder discharge_evaluation

This folder contains the evaluation of model results with discharge data at 1427 discharge stations. Evaluation of modelled discharge included the computation of NSE, KGE and its components as well as streamflow indicator (Q1, Q10, Q25, Q50, Q75, Q90, Q99). Furthermore, plots of all streamflow indicators and efficiency metrics can be found here.

Subfolder twsa

This folder contains all files provided by JPL and CSR, the scripts to produce TWSA series from all four model experiments and the scripts to extract and process GRACE TWSA series. Evaluation of modelled TWSA included R^2, bR^2, γKGE as well as trends in TWSA series. Furthermore, plots of all efficiency metrics can be found here.

Subfolder WBC

This folder contains the script to compute water balance components of all four model experiments as well as the results for two time periods, 1979-2019 and 1981-2010.