Tradeoffs between temporal and spatial pattern calibration and their impacts on robustness and transferability of hydrologic model parameters to ungauged basins

Mehmet Cüneyd Demirel[1], Julian Koch[2], Oldrich Rakovec[3,4], Rohini Kumar[3], Juliane Mai[3,5,6], Sebastian Müller[3], Stephan Thober[3], Luis Samaniego[3], and Simon Stisen[2]

[1]Department of Civil Engineering, Istanbul Technical University
[2]Geological Survey of Denmark and Greenland
[3]Department Computational Hydrosystems, UFZ-Helmholtz Centre for Environmental
[4]Faculty of Environmental Sciences, Czech University of Life
[5]Department of Civil and Environmental Engineering, University of Waterloo
[6]Center for Scalable Data Analytics and Artificial Intelligence-ScaDS.AI

August 24, 2023

Abstract

Optimization of spatially consistent parameter fields is believed to increase the robustness of parameter estimation and its transferability to ungauged basins. The current paper extends previous multi-objective and transferability studies by exploring the value of both multi-basin and spatial pattern calibration of distributed hydrologic models as compared to single-basin and single-objective model calibrations, with respect to tradeoffs, performance and transferability. The mesoscale Hydrological Model (mHM) is used across six large central European basins. Model simulations are evaluated against daily streamflow observations at the basin outlets and remotely sensed evapotranspiration patterns obtained with a two-source energy balance approach. Several model validation experiments are performed through combinations of single- (discharge) and multi-objective (discharge and spatial evapotranspiration patterns) calibrations with holdout experiments saving alternating basins for model evaluation. The study shows that there are very minimal tradeoffs between spatial and temporal performance objectives and that a joint calibration of multiple basins using multiple objective functions provides the most robust estimations of parameter fields that perform better when transferred to ungauged basins. The study indicates that particularly the multi-basin calibration approach is key for robust parametrizations, and that the addition of an objective function tailored for matching spatial patterns of ET fields alters the spatial parameter fields while significantly improving the spatial pattern performance without any tradeoffs with discharge performance. In light of model equifinality, the minimal tradeoff between spatial and temporal performance shows that adding spatial pattern evaluation to the traditional temporal evaluation of hydrological models can assist in identifying optimal parameter sets.

Tradeoffs between temporal and spatial pattern calibration and their impacts on robustness and transferability of hydrologic model parameters to ungauged basins

M.C. Demirel[1,5], J. Koch[2], O. Rakovec[3,8], R. Kumar[3], J. Mai[3,4,6], S. Müller[3], S. Thober[3], L. Samaniego[3,7] and S. Stisen[2]

[1] Department of Civil Engineering, Istanbul Technical University, 34467 Maslak, Istanbul, Turkey

[2]Geological Survey of Denmark and Greenland, Øster Voldgade 10, 1350 Copenhagen, Denmark

[3]Department Computational Hydrosystems, UFZ—Helmholtz Centre for Environmental Research, 04318 Leipzig, Germany

[4]Department of Civil and Environmental Engineering, University of Waterloo, 200 University Ave West, Waterloo ON, N2L 3G1, Canada

[5]Istanbul Technical University, Hydraulics and Marine Sciences Research Center, 34467 Maslak, Istanbul, Turkey.

[6] Center for Scalable Data Analytics and Artificial Intelligence–ScaDS.AI, Humboldtstraße 25, Leipzig, 04105, Saxony, Germany.

[7] Institute of Earth and Environmental Science-Geoecology, University of Potsdam, 14476, Potsdam, Germany

[8] Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol 16500, Czech Republic

Corresponding author: Simon Stisen (sst@geus.dk, ORCID 0000-0001-6695-8412)

Key Points:

- Clear improvements in simulated spatial patterns can be achieved with very limited tradeoff with discharge performance
- Multi-basin and spatial pattern-based calibration improve parameter realism and transferability to ungauged basins

Abstract

Optimization of spatially consistent parameter fields is believed to increase the robustness of parameter estimation and its transferability to ungauged basins. The current paper extends previous multi-objective and transferability studies by exploring the value of both multi-basin and spatial pattern calibration of distributed hydrologic models as compared to single-basin and single-objective model calibrations, with respect to tradeoffs, performance and transferability. The mesoscale Hydrological Model (mHM) is used across six large central European basins. Model simulations are evaluated against daily streamflow observations at the basin outlets and remotely sensed evapotranspiration patterns obtained with a two-source energy balance approach. Several model validation experiments are performed through combinations of single- (discharge) and multi-objective (discharge and spatial evapotranspiration patterns) calibrations with holdout experiments saving alternating basins for model evaluation. The study shows that there are very minimal tradeoffs between spatial and temporal performance objectives and that a joint calibration of multiple basins using multiple objective functions provides the most robust estimations of parameter fields that perform better when transferred to ungauged basins. The study indicates that particularly the multi-basin calibration approach is key for robust parametrizations, and that the addition of an objective function tailored for matching spatial patterns of ET fields alters the spatial parameter fields while significantly improving the spatial pattern performance without any tradeoffs with discharge performance. In light of model equifinality, the minimal tradeoff between spatial and temporal performance shows that adding spatial pattern evaluation to the traditional temporal evaluation of hydrological models can assist in identifying optimal parameter sets.

Plain Language Summary

Hydrological models typically require local observations of river flow to calibrate the models and test their predictive capability. This limits the possibility for predictions in ungauged basins. This study used holdout tests to investigate the robustness of hydrological predictions for ungauged basins. Particularly we investigate how adding more basins and observed spatial patterns of evapotranspiration in the calibration of these models impact this robustness and transferability of model parameters to basins not used for calibration. Results show that transferability and spatial consistency of parameters increase when adding more basins and spatial pattern observations.

1 Introduction

High-resolution distributed hydrological models are increasingly being employed to address and solve a broad range of water-related issues (Bierkens et al., 2015). Output from such models is used to analyze hydrological responses and climate impact assessments at a scale far below the spatial scale at which the models are calibrated and validated (Mizukami et al., 2017; Samaniego et al., 2017). Similarly, models are generalized based on parameterizations obtained from neighboring basins, which is particularly pertinent for efficiently managing ungauged basins (Hrachowitz et al., 2013). This poses two fundamental challenges; how do we improve the reliability of model simulations at a higher spatial resolution and how do we develop robust parametrizations that can be transferred meaningfully to neighboring catchments? (Fenicia et al., 2014; Kirchner, 2006; Kumar, Samaniego, et al., 2013; Samaniego et al., 2010).

Integration of satellite remote sensing data with distributed hydrological models has been a common path towards improving the reliability of hydrological model simulations (Dembélé et al., 2020). This development has followed with the progress and availability of remotely sensed datasets, which have evolved significantly over the past decades, although accuracy of satellite-based datasets remains varying (Ko et al., 2019; Stisen et al., 2021).

Several studies have addressed the impacts of adding remotely sensed observations to streamflow calibration (Odusanya et al., 2022; Rientjes et al., 2013; Sirisena et al., 2020). Nijzink et al. (2018) presented a large modelling effort illustrating the impact of adding different remotely sensed products across five different conceptual model and 27 European catchments. They analyzed 1023 possible model combinations regarding model constraint and showed an added value of remotely sensed data in the absence of streamflow data. In a recent model intercomparison paper Mei et al. (2023) analyzed different model calibration strategies combining streamflow and global gridded soil moisture and evapotranspiration datasets. They found that adding soil moisture to the streamflow calibration improved evapotranspiration performances. Mei et al. (2023) also included a review of 16 previous papers on the subject of constraining models using a combination of streamflow and remotely sensed data. Both the study by Nijzink et al. and Mei et al., and 14 out of the 16 papers in Mei et al. review applied spatially averaged time series of the remotely sensed data. By this approach, the spatial information in the satellite data is ignored and the hydrological model evaluation remains limited to the temporal component of the models. This traditional focus on the temporal performance is due to several factors related to either the spatial resolution of the models and the remote sensing data or lack of spatial performance metrics and optimization frameworks.

Therefore, an particular avenue addressing the challenges of distributed model fidelity, is the use of spatial pattern information from remote sensing data to constrain hydrological model parametrization (Dembélé et al., 2020; Demirel, Mai, et al., 2018; Koch et al., 2022; Soltani, Bjerre, et al., 2021; Zink et al., 2018).

The fundamental idea behind this approach is to employ a multi-objective calibration framework that adds to the traditional discharge-based calibration, an independent set of objective functions that mainly reflects the observed spatial pattern of key hydrological states or fluxes. This approach differs from multi-objective calibrations based on multiple metrics calculated from the same observation (e.g. streamflow timeseries) or application of basin average timeseries of remotely sensed data (Demirel, Mai, et al., 2018). In addition, independence in the optimization approach can be obtained by adding the new information source in combination with a pareto-achieving optimizer which circumvents the need to join multiple objective functions into a single score (Mei et al., 2023).

A previous study by Zink et al. (2018) incorporated land surface temperature patterns in model calibration and showed that this helped to better constrain the model parameters connected to evapotranspiration when compared to calibrations based on streamflow only. Moreover, in their study the model performance regarding evapotranspiration increased at seven eddy flux measurement sites used for evaluation. Adding new constraints to calibration decreased streamflow performance yet the authors of that study illustrated how land surface temperature data could secure better results for ungauged basins. For a single Danish basin, Demirel, Mai, et al. (2018) developed a spatial pattern-oriented calibration framework and a new spatial performance metrics, and illustrated a small tradeoff between streamflow and spatial pattern performance. Dembélé et al. (2020) applied a similar calibration framework to a model study of the poorly gauged Volta River basin in West Africa. They showed that while streamflow and terrestrial water storage performance decreased by 7% and 6 %,

respectively, soil moisture and evapotranspiration performances increased by 105% and 26% respectively when including the spatial calibration framework with multiple objectives. Soltani et al. (2021) illustrated how adding spatial pattern optimization to a national scale groundwater model improved evapotranspiration patterns and altered groundwater recharge patterns without deteriorating groundwater head and discharge performance significantly. Other, recent studies such as Xiao et al. (2022) and Ko et al. (2019) have utilized spatial patterns of land surface temperature for hydrological model evaluation. However, in the context of our current study, Xiao et al. (2022) and Ko et al. (2019) did not address the tradeoffs between different optimization strategies and streamflow performance.

As a results of increased availability of remotely sensed datasets combined with machine learning approaches and computational power, many gridded spatial products are now available (Belgiu & Drăguţ, 2016; Feigl et al., 2022). These all facilitate the spatial characterization of hydrologic variables and fluxes and enable spatial model evaluations. However, to optimize the simulated spatial patterns of a hydrological model, the model parametrization scheme needs to be fully distributed and spatially flexible. In this context, the multi-scale parameter regionalization (MPR) method (Samaniego et al., 2010) represented a significant advancement, which was initially included in the mesoscale Hydrological Model (mHM) (Kumar, Samaniego, et al., 2013; Samaniego et al., 2010). Afterwards, it has been incorporated into several other modelling frameworks (Lane et al., 2021; Mizukami et al., 2017; Tangdamrongsub et al., 2017) and it is available as a stand-alone parametrization tool that can be coupled to hydrological models (Schweppe et al., 2022). Other studies have developed similar flexible parametrizations schemes based on pedo-transfer functions using gridded data (Feigl et al., 2020; Ko et al., 2019).

It is well known that streamflow calibration does not guarantee good spatial pattern performance (Rakovec, Kumar, Mai, et al., 2016; Stisen et al., 2011) and performance on the initial single objective typically drops when adding additional objectives to the calibration. But what is the tradeoff between spatial and temporal performance? How does single and multi-objective optimization impact parameter transferability, and how does this compare to impacts of multi-basin optimization? Based on the above, we aim at addressing the following research gaps:

- What are the tradeoffs between temporal and spatial model performance investigated in a pareto-achieving optimization framework?
- How does multi-basin and spatial pattern-oriented calibration impact model performance and transferability to ungauged basins?

In this study, we demonstrate the impact of multi-site and multi-objective calibration compared to single-site and single-objective parameter estimation, i.e., the most common practice in hydrologic modeling, specifically in the context of parameter transferability to ungauged basins. The impact on parameters transferability via adding spatial patterns into the model calibration, is a novel aspect that has not received much attention in the literature so far. The study is conducted for six mesoscale central European basins.

The distributed modelling study is carried out in the framework of a flexible spatial model parameterization scheme in combination with observed spatial patterns of actual evapotranspiration (AET) derived from satellite data. We apply the mHM model code since it suit the applied calibration framework well due to its flexible model parametrization schemes based on pedo-transfer functions to distribute soil parameters and the built-in multi-scale parameter regionalization.

We design a set of model calibration experiments including both single- and multiple basins as well as single- and multi-objective calibrations and two jack-knife experiments, i.e., sequentially keeping one or five of six basins out of the joint calibration approach. Model simulations are evaluated based on temporal discharge performance and spatial AET performance using long term average monthly pattern maps, appropriate objective functions and a global multi-objective pareto-achieving search algorithm is applied to illustrate the exact tradeoff between the two objectives.

2 Methodology

Catchments, observed data (both Section 2.1), and the hydrologic model (Section 2.2) are presented in this section. The objective functions to evaluate the model performance of simulated discharge and AET, respectively, are described in Section 2.3. A sensitivity analysis performed to determine the most important parameters for model calibration is described in Section 2.4, while Section 2.5 describes the calibration and validation setup including a brief description of the multi-objective calibration algorithm applied.

2.1 Catchments and hydro-meteorological data

This study is conducted using six European catchments, i.e., Elbe, Main, Meuse, Mosel, Neckar, and Vienne with drainage areas varying from 12,775 km^2 to 95,042 km^2. The catchments are spread over Central Europe and represent a diversity of soil texture, land use, and land cover. The mean annual rainfall varies from 637 mm to 874 mm while the mean annual runoff varies from 184 mm to 398 mm (Table 1). The six catchments are selected based on two criteria: good model performance obtained in previous studies (Rakovec, Kumar, Mai, et al., 2016) and spatial patterns of AET that are likely dominated by land-surface heterogeneity, i.e., land cover and soil properties, rather than a strong climate gradient. The latter will facilitate a meaningful model calibration driven by spatial patterns since simulated patterns can be adjusted through the surface parametrization within the hydrological model and are not purely driven by climate (Koch et al., 2022). In basins with a large climate gradient, simulated spatial patterns are typically easier to simulate even with a suboptimal spatial parametrization since the patterns are to a lesser degree controlled by the model parameters and will display correct overall patterns enforced by the climate forcing data.

Average temperature, precipitation, and PET data are available at daily time steps and over 0.25-degree grids for the period extending from 1980 to 2018, whereas the length of the observed daily discharge data varies between catchments. Daily averaged meteorological data (P and PET) were obtained from the E-OBS and ERA-5 reanalysis datasets (Cornes et al., 2018; Hersbach et al., 2020). PET was estimated based on the Hargreaves–Samani model using ERA-5 air temperature data (daily minimum, maximum, and mean) as input (George H. Hargreaves & Zohrab A. Samani, 1985).

In addition to the six outlet discharge gauges used in model calibration, we obtained daily data from 46 gauging stations from the Global Runoff Data Center (GRDC, 2020) for internal validation of the six catchment models. Remotely sensed AET estimates for the period from 2002 to 2014 were obtained using MODIS data and the two-source energy balance method (Norman et al., 1995) as described in Stisen et al. (2021). Digital elevation model (DEM) data were retrieved from the Shuttle Radar Topographic Mission (SRTM, Farr et al., 2007). Soil texture variables, clay content, sand content, and bulk density were derived from the SoilGrid Database (Rakovec et al., 2019; Rakovec, Kumar, Attinger, et al., 2016). The soil texture data for six layers with varying depths (5, 15, 30, 50, 100, and 200 cm) and a tillage depth of 30 cm are introduced as input to the model. All input data were resampled to

a common spatial resolution of 0.001953125 degree (~200 m) (Rakovec, Kumar, Mai, et al., 2016). MODIS-based land use was reclassified into three classes, namely forest, pervious, and impervious. Long-term monthly LAI maps used to calculate the spatiotemporally varying crop coefficient are based on the MODIS MOD16A2.v061 product (Running et al., 2021). The original eight-day composite LAI maps were aggregated to long-term monthly means at a matching spatial resolution of ~200 m.

Table 1 Main characteristics of the six catchments, i.e., drainage area (km^2), annual precipitation (P in mm), annual potential evapotranspiration (PET in mm), and annual discharge (Q in mm) calculated based on the common period of 1980-2016.

	Area (km^2)	P (mm)	PET (mm)	Q (mm)
Elbe	95,042	637	755	184
Main	14,117	736	773	243
Meuse	20,143	874	741	398
Mosel	27,127	872	777	365
Neckar	12,775	858	782	335
Vienne	19,892	815	864	308

2.2 Hydrologic model

The spatially explicit mesoscale Hydrologic Model v.5.11.1 (Kumar, Samaniego, et al., 2013; Rakovec et al., 2019; Samaniego et al., 2010; Thober et al., 2019) was used to simulate daily discharge and spatial AET patterns of the six catchments. The backbone of mHM, i.e. the numerical methods utilized to estimate various states and fluxes are based on the fusion of two well-known models, i.e., HBV and VIC (Samaniego et al., 2010). mHM simulates major components of the hydrologic cycle, i.e., evapotranspiration, canopy interception, snow accumulation and melting, soil moisture dynamics, infiltration, percolation, groundwater storage, and surface runoff generation. The model simulates these fluxes on a multi-layer distributed grid using the multi-scale parameter regionalization approach (Kumar, Samaniego, et al., 2013; Samaniego et al., 2010) to account for sub-grid variability of landscape attributes and model parameters based on pedo-transfer functions. MPR is one of the unique features of mHM that facilitates a spatial pattern-oriented calibration. Moreover, AET and soil moisture from different soil layers are modelled based on available soil water and the root fraction of vegetation in each soil layer. Two transfer functions are of particular importance for this work. Firstly, an exponential function Eq. (1) uses monthly LAI maps to link distributed vegetation dynamics to a spatially distributed crop coefficient, also termed a dynamic scaling function (Demirel, Mai, et al., 2018). The spatially distributed crop coefficient is then applied for scaling spatially coarse PET data to account for a heterogenous land cover.

$$K_c[m] = K_{c,min} + \left(K_{c,min} + K_{c,max}\right)\left(1 - e^{a*LAI[m]}\right) \quad (1)$$

Secondly, another transfer function utilizes spatially distributed soil texture maps to allow for an incorporation of soil physical properties in the spatial parametrization of root

fraction coefficients (Demirel, Mai, et al., 2018). Hereby, the root fraction coefficient can vary with both vegetation and soil type and is used in mHM to calculate root water uptake as part of the AET reductions factor (Samaniego et al., 2021). During calibration, both transfer functions increase the model flexibility to adjust the spatial AET patterns retrieved from satellite data (Demirel, Mai, et al., 2018; Koch et al., 2022). Finally, the total runoff generated at every grid cell is routed to its neighboring downstream cell using the adaptive timestep spatially varying celerity method for the river runoff routing scheme (Thober et al., 2019).

mHM has previously been parameterized and successfully calibrated against multiple satellite-based datasets including AET and terrestrial water storage anomalies (GRACE), land surface temperature, and soil moisture at multiple spatial scales over numerous river basins (Busari et al., 2021; Ekmekcioğlu et al., 2022; Koch et al., 2022; Kumar, Livneh, et al., 2013; Rakovec, Kumar, Attinger, et al., 2016; Rakovec, Kumar, Mai, et al., 2016; Zink et al., 2018).

In this study, the following four different spatial resolutions are defined in the mHM model: 0.001953125 degree for the morphological characteristics (L0 scale), 0.015625 degree for the hydrologic modeling resolution (L1), 0.0625 degree for runoff routing (L11), and 0.25 degree for the meteorological forcing (L2). Note that around 50 degrees North these resolutions correspond approximately to 140/430 m, 1.1/3.5 km, 4.5/7 km and 18/28 km lon/lat for L0, L1, L11 and L2 respectively. Finally, a 13-year period (2002–2014) with a 4-year warming period (1998-2001) was simulated at a daily timestep for calibration and evaluation of the discharge performance and spatial pattern match between remote sensing based and simulated AET (from 2002 to 2014). The remote sensing based AET is estimated with the Two-Source Energy Balance method (TSEB) (Norman et al., 1995), using MODIS data including land surface temperature, albedo and NDVI. For a full description of the AET dataset and comparison to other estimates for Europe the reader is referred to (Stisen et al., 2021).

2.3 Evaluation metrics and objective functions

The hydrologic model performance was evaluated using two key objective functions (OFs). In this study, we interpret multi-objective calibration as the combination of two completely independent evaluation datasets, i.e., discharge time series and spatial AET maps, instead of producing a variety of OFs based on a single variable. For the temporal evaluation of discharge, the Kling-Gupta-Efficiency (KGE) was applied (Gupta et al., 2009). The KGE is defined as

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

where $\boldsymbol{r}$ is the Pearson correlation coefficient between observed and simulated streamflow (Q), $\boldsymbol{\alpha}$ is the variability which is defined as the ratio of the standard deviation of observed and simulated Q, and $\boldsymbol{\beta}$ is the bias defined as the ratio between average observed and simulated Q.

For the spatial pattern evaluation of simulated AET, the bias-insensitive Spatial Efficiency metric (SPAEF) was used (Demirel, Mai, et al., 2018; Koch et al., 2018). The SPAEF is a reformulation of KGE and is defined as

$$SPAEF = 1 - \sqrt{(r-1)^2 + (\Phi-1)^2 + (\gamma-1)^2} \quad (3)$$

where $\boldsymbol{r}$ is the Pearson correlation coefficient between observed and simulated spatial patterns of AET, $\boldsymbol{\Phi}$ is the coefficient of determination fraction of observed and simulated

AET, and γ quantifies the fraction of the histogram intersection based on the z-scores of observed and simulated AET fields. Both, KGE and SPAEF vary in a range from -∞ to the best value of 1.

SPAEF is calculated separately for long-term seasonal averages across all years, focusing on the water limited growing seasons in three three-month windows i.e. March-April-May (MAM), June-July-August (JJA), and September-October-November (SON). The fourth quarter (December-January-February) is not used because there are cloud issues in winter and energy limited conditions dominate spatial patterns of AET. Finally, we summed the squared residual of these three seasonal parts as

$$SSR_{AET} = (1 - SPAEF_{MAM})^2 + (1 - SPAEF_{JJA})^2 + (1 - SPAEF_{SON})^2 \quad (4)$$

where SSR_{AET} represents the sum of squared residuals for the seasonal AET pattern performance applying SPAEF as OF. For the joint calibrations across several catchments SPAEF is calculated on the combined dataset as a single value for each season across all catchments.

For streamflow, KGE was calculated at one, five or six stations (n) at the outlet of the basins and the sum of squared residuals is used

$$SSR_Q = \sum_i^n (1 - KGE_i)^2 \quad (5)$$

where SSR_Q represents the sum of squared residuals for the Q performance at the discharge stations using KGE as OF. For the single catchment calibrations, only the corresponding station KGE is utilized.

2.4 Sensitivity Analysis

Identification of the optimal parameter set through a calibration framework can be cumbersome if the dimension of the search space is not limited by a sensitivity analysis first. mHM has 69 parameters (Samaniego et al., 2021) each increasing the dimension of the search space. Focusing a calibration on only parameters that are sensitive regarding the selected OFs is computationally more efficient than calibrating all parameters (Demirel, Koch, et al., 2018). To reduce the computational burden by narrowing the search space, a one-at-a-time (OAT) sensitivity analysis was conducted to identify the most important parameters for calibration using the PEST Toolbox (Doherty, 2010). Although the parameter interactions are not accounted for in this local OAT method, it provides an indication of sensitive parameters especially if combined with the expert opinion which can complement the assessment of parameter interactions. We used KGE as OF for discharge performance and SPAEF as OF for spatial pattern performance of the model. Also, geo-parameters and root-transfer function parameters of mHM were analyzed separately for deeper assessment. Each parameter was perturbated two times (5% increased and 5% decreased based on the initial point) to calculate the average sensitivity index of OFs for the change in the parameter value. This index value is then multiplied by the absolute parameter value to account for the parameter magnitude in the calculations. Finally, the sensitivities are normalized by the maximum of the group.

2.5 Experimental design of calibration and validation

In total, 26 calibration experiments were designed to investigate the potential benefits of incorporating AET to augment a multi-objective and multi-basin calibration framework (**Figure *1***). Note that SSR_Q is incorporated in all calibration experiments as objective

function while SSR_{AET} is used only for the KSP1, KSP5 and KSP6 calibration experiments. Note that KSP stands for KGE and SPAEF multi-OF calibration whereas KGE stands for KGE only based single-OF calibration. The indices 1, 5 and 6 show the number of basins included in the calibration experiments as conceptualized in **Figure *1***. In this study, we did not include an AET-only scenario as it failed to reproduce reasonable water balances in our preliminary tests and also a previous study (Demirel, Mai, et al., 2018).

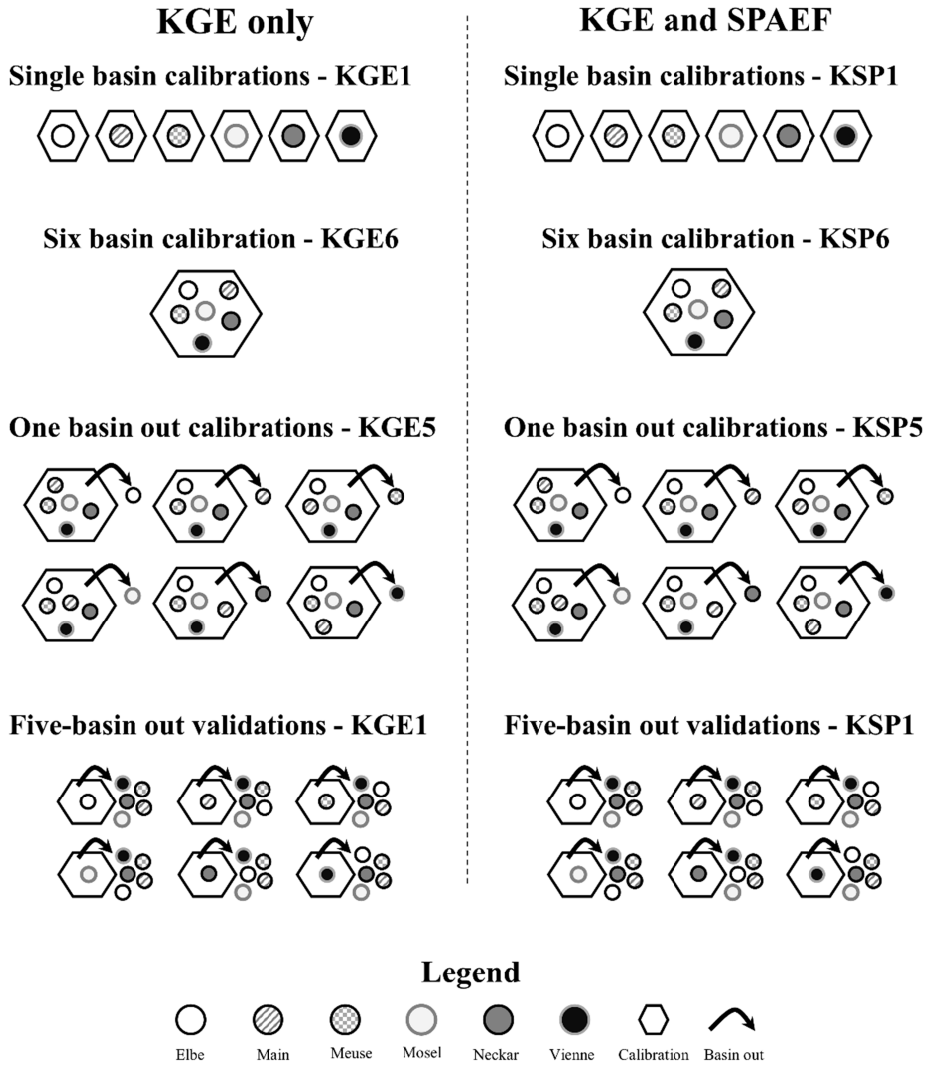


Figure 1: Calibration framework. Six calibration experiments are performed: three using only streamflow observations (left column of panels) using KGE as performance metric and three experiments using streamflow and AET patterns (right column of panels) with two objective functions (KGE and SPAEF) in a KGE-SPAEF-Pareto (KSP) experiment. The six basins are either calibrated independently (first row), or collectively (second row), or in a one-basin-out at a time, i.e., Jack-knife robustness test approach (third row).

All 26 calibration experiments (cases) were performed with the open-source model-agnostic Ostrich optimization toolbox written in C++ (Matott, 2017). For all 26 calibration experiments, the parallel implementation of the Pareto-Archived Dynamically Dimensioned Search (ParaPADDS) algorithm was used (Asadzadeh & Tolson, 2013). This algorithm is the multi-objective version of the Dynamically Dimension Search (Tolson & Shoemaker, 2007) algorithm that identifies a Pareto front of non-dominated optimal solutions, which is most appropriate for our multi-objective calibrations (Beume & Rudolph, 2006; Razavi & Tolson, 2013). Moreover, ParaPADDS algorithm reached reasonable solutions for both single and multiple OFs; therefore, we used the same search algorithm in all scenarios for consistency. The ParaPADDS algorithm was configured with user-defined maximum 750 iterations, with 3 parallel nodes (logical processors), a perturbation value of 0.2, and the exact hypervolume contribution as the selection criterion. Note that initial tests for one basin with 200, 500 and 1000 iterations indicated stable results already at 500, but a somewhat incomplete Pareto-front. Based on this and in the interest of saving computation time, we decided on 750 iterations. Like all multi-objective calibration methods, the algorithm does not provide a single best solution for the multiple OF problem. Still, it offers the modeler a set of possible solutions on the Pareto front (Asadzadeh & Tolson, 2013).

KGE1 and KGE6 calibrations resulted in the single best parameter set that was used to create our final results in the following figures. KSP1 and KSP6 calibrations provided multiple possible solutions on the Pareto front with KGE as one axis and SPAEF as the other axis. To systematically select a best-balanced parameter set, we picked the solution that is closest to the origin by normalizing both axes (SSR_Q and SSR_{AET}) using min-max normalization and choosing the minimum of the sums, similar to the approach by Martinsen et al. (2022). The normalization is applied to avoid the metric-magnitude effects on the selection. KSP1 and KSP6 results presented hereafter are generated using this selected single parameter set. Calibrations were done with the six discharge gauges and three seasonal AET maps (March to November). We used 46 discharge stations from GRDC for internal validation of the six catchment models and we show the results of KGE5 and KSP5 cases as maps (See Figure 6).

3 Results

3.1 Sensitivity Analysis

Table 2 shows the 20 most influential parameters out of 69 mHM parameters selected based on the combined sensitivity of the two metrics. We used these normalized sensitivities varying from 0 to 100% and applied a threshold of 1% for at least one of the OF's for selecting the most sensitive parameters for calibration (20). Based on the KGE, the five most sensitive parameters controlling discharge are RotFrCofClay, RotFrCofFore, PTFLowConst, PET_apervi, PTFKsConst which are parameters mainly controlling the AET and thereby the water balance. KGE is also sensitive to some routing parameters but generally less than the parameters controlling AET levels. The SPAEF OF is most sensitive to the parameters RotFrCofClay, RotFrCofFore, PET_apervi and PET_aforest, which is almost identical to the most sensitive parameters for KGE. Additionally, parameters associated with simulated patterns, e.g., related to pedo-transfer functions for soil properties are important for SPAEF. Conversely, SPAEF has zero sensitivity to routing parameters. Overall, the most sensitive parameters contribute to spatial heterogeneity of root fraction coefficients, crop coefficients, infiltration factor and field capacities, of the grid cells.

Table 2: 20 selected parameters for calibration, their range and sensitivity for both objective functions KGE and SPAEF. The parameter abbreviations correspond to the name of the parameter in the mHM setup (In the mHM namelist).

Parameter abbreviation (in the mHM namelist)	**Description**	**Range**	**Normalized Sensitivity [%]**	
			KGE	**SPAEF**
ExpSlwIntFlW	Exponent slow interflow	0.05 to 0.3	3.6%	0.0%
InfShapeF	Infiltration shape factor	1 to 4	2.6%	1.2%
IntRecesSlp	Interflow recession slope	0 to 10	1.2%	0.0%
PET_aforest	Intercept – forest in dynamic scaling function (K_c) for PET	0.3 to 1.3	10.1%	19.1%
PET_aimpervi	Intercept – impervious in dynamic scaling function (K_c) for PET	0.3 to 1.3	0.5%	2.7%
PET_apervi	Intercept – pervious in dynamic scaling function (K_c) for PET	0.3 to 1.3	15.3%	15.2%
PET_bb	Base coefficient for K_c	0 to 1.5	3.7%	2.2%
PET_cc	Exponent coefficient for K_c	-2 to 0	1.6%	0.9%
PTFHigConst	Constant in Pedo-transfer function for soils with sand content higher than 66.5%	0.5358 to 1.1232	0.3%	1.1%
PTFKsconst	Constant in pedo-transfer function for hydraulic conductivity of soils with sand content higher than 66.5%	-1.2 to -0.285	11.7%	0.5%
PTFKssand	Coefficient for sand content in pedo-transfer function for hydraulic conductivity	0.006 to 0.026	3.5%	0.3%
PTFLowclay	Constant in Pedo-transfer function for soils with clay content lower than 66.5%	0.0001 to 0.0029	1.5%	1.5%
PTFLowConst	Constant in Pedo-transfer function for soils with sand content lower than 66.5%	0.6462 to 0.9506	21.1%	13.9%
PTFLowDb	Coefficient for bulk density in Pedo-transfer function for soils with sand content lower than 66.5%	-0.3727 to -0.1871	10.9%	8.7%
RechargCoef	Recharge coefficient	0 to 50	1.9%	0.0%
RotFrCofClay	Root fraction for clay	0.9 to 0.999	100.0%	100.0%
RotFrCofFore	Root fraction for forest areas	0.9 to 0.999	57.2%	75.9%
RotFrCofImp	Root fraction for impervious areas	0.9 to 0.999	2.1%	9.3%
RotFrCofSand	Root fraction for sand	0.001 to 0.09	1.9%	2.3%
SlwIntReceKs	Slow interception	1 to 30	1.4%	0.0%

3.2 Calibration Results

Figure 2 shows the model calibration results for single basin calibrations using single (KGE1) and multi-objective functions (KSP1). Each calibration is performed using 750 model runs distributed in three parallel processors where non-dominated runs leading to a Pareto front are identified by the ParaPADDS algorithm (Asadzadeh & Tolson, 2013). A solution is called non-dominated if there is no other solution that is better in all objectives analyzed. Although the calibrations do not depict a clear Pareto front due to the combined plotting of KSP1 and KGE1, the tradeoff between only discharge and spatial performances is clearly distinguishable from the plots. Only KGE (KGE1) calibrations lead to slightly better KGE performance and much poorer spatial AET pattern performance than those in multi-objective calibrations (KSP1). However, KSP1 calibrations enable to identify a more-balanced solution leading to higherSPAEF performance and only slightly poorer KGE performance than KGE1's single solution for each basin (shown as a triangle). While it is well known that good KGE performance does not guarantee a good spatial pattern performance, it is a novel finding that there is a very limited tradeoff between the temporal and spatial performance of the models.

Table 3 shows the KGE performance for the best-balanced solutions from KGE1 and KSP1 along with the SPAEF calculated across all six basins when combining the six single basin calibrations. Generally, all calibrations can lead to KGE performances in the range of 0.84 to 0.96 as shown in Table 3.

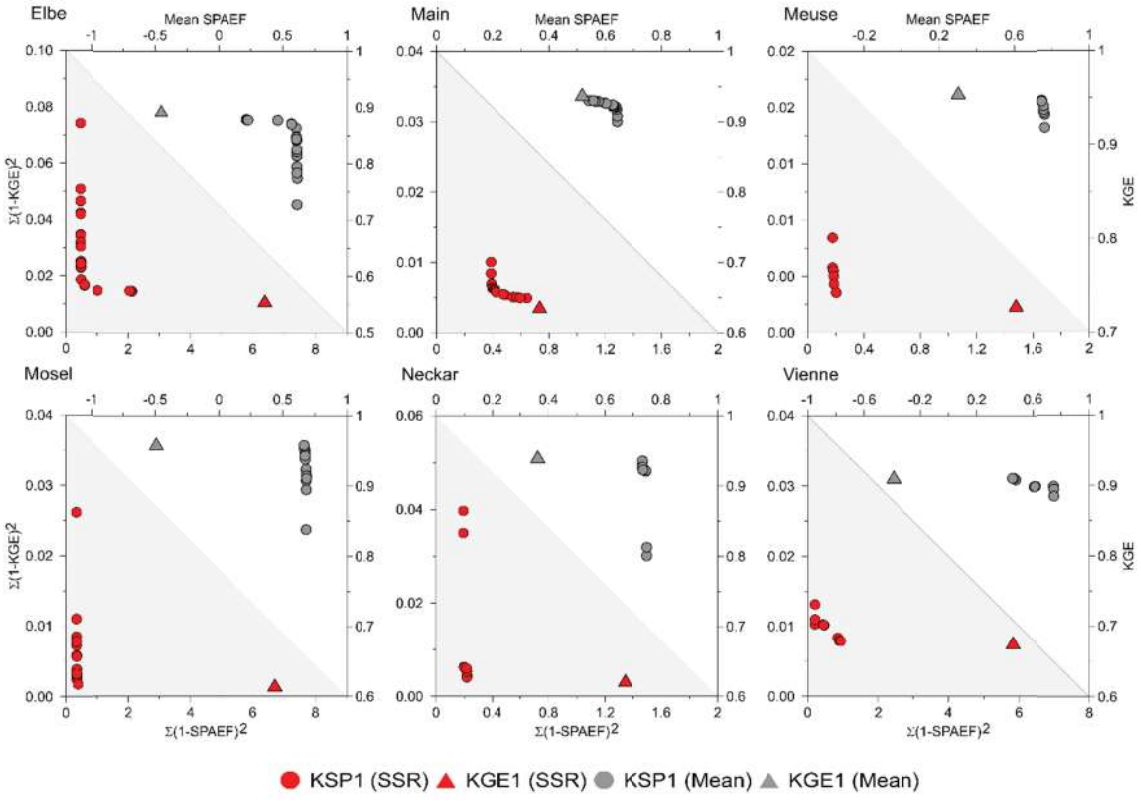


Figure 2: Calibration results for six basins. Note that red symbols in the grey zone (lower left) are the exact OF values (sum of squared errors) used for calibration, whereas grey

symbols in the white zone (upper right) are the corresponding metric values. We plot both as it is easier to relate and compare actual metric values. Mean SPAEF values refer to the mean of SPAEF calculated for the three seasons.

Subsequently, a multi-basin calibration was conducted again with both single (KGE6) and multiple (KSP6) objectives (see Method Section 2.5 for details). The results are shown in Figure 3 and Table 3. The model performance results mimic the results of the single basin test, with similar KGE performances; however, with a significant performance increase for SPAEF, from 0.02 with KGE6 to 0.61 with KSP6, as would be expected when adding the spatial pattern objective function. Table 3 highlights the limited tradeoff for KGE both for individual stations and averages.

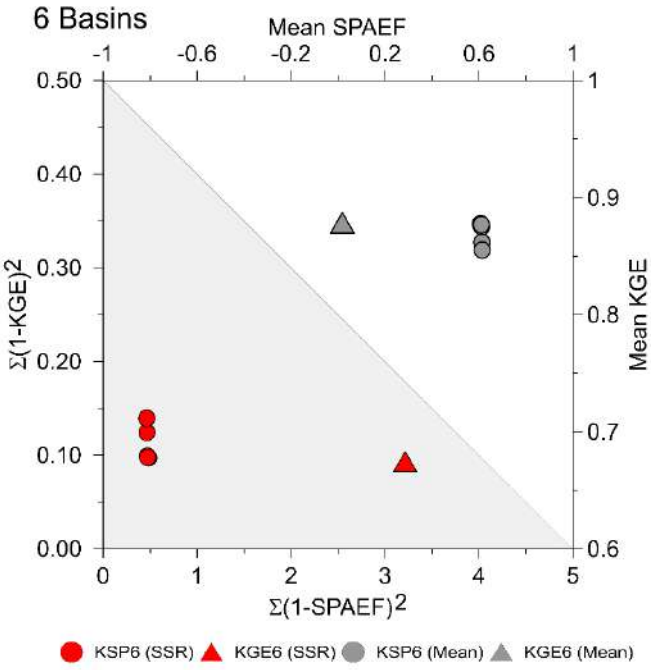


Figure 3: Multi-basin calibration results across the six basins for KGE only (triangles) and for KGE and SPAEF (circles) as objective functions. Note that the grey zone (lower left panel) is the exact OF values used for calibration whereas the white zone (upper right panel) is the corresponding metric value.

Figure 4 illustrates the spatial AET maps from TSEB (observed) and the various calibration tests. For the multi-objective calibrations (KSP1 and KSP6), the best-balanced solution (closest point to the origin) is chosen for visualization. The maps clearly show the issues related to KGE1, regarding spatial pattern performance. For three out of six basins, i.e., Elbe, Mosel and Vienne, the KGE1 calibration has resulted in a strikingly poor spatial AET pattern (compared to KSP1) where distinct low and high AET areas were inverted as compared to the TSEB pattern. In contrast, including the SPAEF metric in the optimization (KSP1) prevented such errors without any substantial loss in KGE performance (average KGE of 0.93 for KGE1 and 0.90 for KSP1, Table 3).

Interestingly, the KGE6 calibration, i.e., without any spatial pattern constraint, was able to represent the overall pattern to some extent across the six basins, although with a significantly underestimated variance and some substantial differences. This emphasizes the value of joint multi-basin calibration for robustness in spatial parametrization within the MPR

parametrization scheme. Adding the SPAEF metric to the multi-basin calibration (KSP6), generated the best spatial similarity to TSEB, although not better than combining spatial AET from the six individual KSP1 calibrations maps into one map (Figure 4 and Table 3). Comparing KGE1 and KGE6 calibrations illustrates the reduction in KGE performance from averages of 0.93 falling to 0.88, when seeking one common parametrization in KGE6. The higher KGE performance obtained from single basin optimization does however come with a very poor SPAEF performance of -0.45 for KGE1 compared to 0.02 for KGE6. Although the SPAEF for KGE6 is also low, this is mainly attributed to the variance component of SPAEF (Figure 4).

Table 3: Model performances on KGE and SPAEF (across all basins) for different calibrations experiments. Values for the KSP-calibrations represent the best-balanced solutions from the pareto fronts. Values in parentheses are STD across stations for Average KGE and across seasons for SPAEF.

Basin	**KGE1**	**KSP1**	**KGE6**	**KSP6**
Elbe KGE	0.89	0.84	0.87	0.84
Main KGE	0.94	0.91	0.84	0.84
Meuse KGE	0.96	0.93	0.91	0.93
Mosel KGE	0.96	0.91	0.90	0.90
Neckar KGE	0.94	0.92	0.90	0.89
Vienne KGE	0.91	0.90	0.85	0.86
Average KGE	**0.93** (0.02)	**0.90** (0.03)	**0.88** (0.03)	**0.88** (0.04)
Across Basins SPAEF	**-0.45** (0.18)	**0.61** (0.06)	**0.02** (0.40)	**0.61** (0.10)

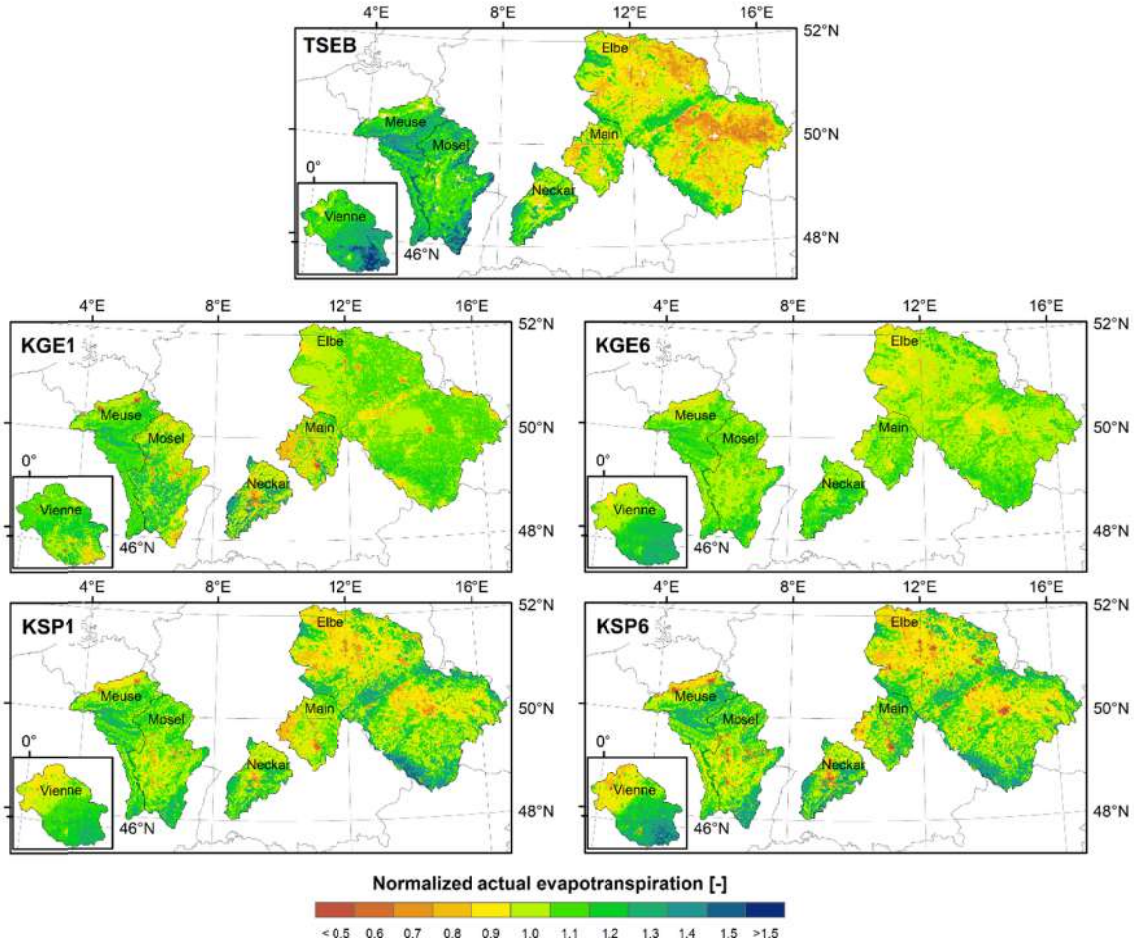


Figure 4: Spatial patterns of normalized AET (average of March to November) from TSEB model (first row), KGE-only calibration cases (second row), multi-objective calibration cases using KGE and SPAEF (third row). The left column shows calibration results when the six basins are calibrated independently while the right column shows the results when the six basins are calibrated collectively.

Even though the model performance of simulated spatial patterns across the six basins shares some similarities for KSP6 and KSP1, there is a marked difference between the parameter distributions that generate the spatial AET patterns. This is shown in Figure 5 displaying the resulting parameter fields of field capacity and crop coefficient, which are calculated in mHM and represents the key controls of AET simulations. The field capacity and crop coefficients are not parameters that are assigned directly in mHM but are the results of several transfer function parameters. Therefore, field capacity and crop coefficient are not included in Table 2, which lists the transfer parameters that generate them (PET* and PTF*). Although KSP1 calibrations generate parameter distributions that have meaningful patterns of field capacity within each basin, it fails to form one consistent seamless parametrization across the basins (Figure 5). Similarly, for the KGE1 and KGE6 calibrations, the spatial inconsistency resulting from single basin calibration becomes apparent. For field capacity (Figure 5), the parametrization obtained from KGE6 and KSP6 is relatively similar, although the KSP6 results in a slightly larger variance. A different picture emerges for the crop coefficient where KSP1 generates patterns similar to KSP6. At the same time KGE6 produces very different

patterns with unreasonably high values for urban areas and little impact of vegetation patterns on crop coefficients. This difference is due to the crop coefficient parameter pattern mainly being constrained by the SPAEF OF, while the KGE OF on discharge also constrains the field capacity parameter.

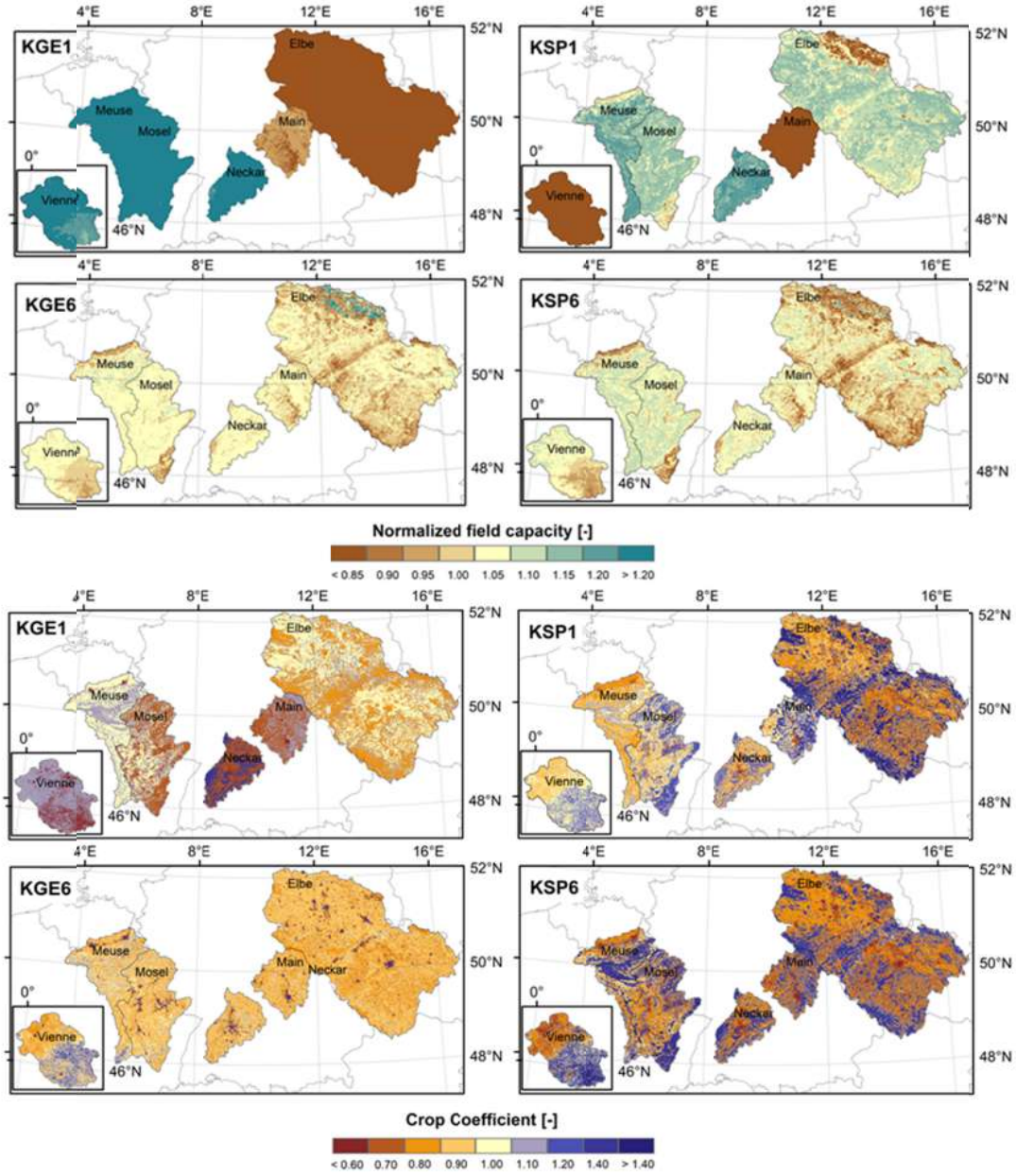


Figure 5: Spatial patterns of estimated normalized field capacity (top four panels) and crop coefficient (bottom four panels) when calibrating only streamflow for each basin independently (KGE1) or collectively (KGE6) or using streamflow and AET to constraint each basin independently (KSP1) or collectively (KSP6). Note that field capacity and crop coefficient are not direct parameters of mHM but are calculated from several parameters (PET* and PTF*) in Table 2.

3.3 Cross-validation Results

To investigate the potential impact of the calibration strategy on the transferability of parameters to ungauged basins, two Jack-knife tests were applied. The two tests are holding out five (KGE1-KSP1) or one (KGE5-KSP5) basins simultaneously and evaluating only the uncalibrated basins using parameters obtained calibrating either one or five other basins. These tests are performed for both single- and multi-objective calibrations, resulting in four parameter transfer tests.

Results for the single-basin calibrations and subsequent evaluation of the performance of parameter transfer to five ungauged basins based on the KGE1 and KSP1 calibrations are shown in Table 4. For each discharge evaluation, KGE is calculated as the average across all basins, each represented in five holdout evaluations (a total of 30 ungauged evaluations). The SPAEF is calculated based on three seasons for six holdouts (a total of 18 pattern evaluations). Table 4 shows that discharge performances with average KGE of 0.79 and 0.83 across ungauged basins, and similar between KGE1 and KSP1, although the latter performs better. Compared to the KGE6 and KSP6 calibrations (both with an average KGE of 0.88 in Table 3, relatively little loss in performance for discharge is noticed, even for ungauged cases.

For the spatial pattern evaluation, the performance for the KGE1 parameter transfer has low average SPAEF across all basins, while the standard deviations are large across seasons. For KSP1, the results of SPAEF are much better with an average of 0.41. This indicates that single basin calibration with multiple objectives can better make robust predictions for ungauged basins when both discharge and AET patterns are considered in calibration at gauged locations.

Table 4: Model performances on KGE and SPAEF (across all basins) for different cross-validation experiments. Values for the KSP-calibrations represent the best-balanced solutions from the Pareto fronts. For KGE1 and KPS1 KGE values are averages across five holdout experiments. Values in parentheses are STD across holdout experiments for single stations and across stations and holdout solutions for average KGE and for SPAEF STD is calculated across seasons and holdout solutions.

Basin	**KGE1 holdout**	**KSP1 holdout**	**KGE5 holdout**	**KSP5 holdout**
Elbe KGE	0.72 (0.05)	0.76 (0.06)	0.83	0.84
Main KGE	0.76 (0.09)	0.77 (0.04)	0.81	0.80
Meuse KGE	0.84 (0.08)	0.91 (0.03)	0.89	0.94
Mosel KGE	0.85 (0.03)	0.88 (0.04)	0.88	0.87
Neckar KGE	0.82 (0.08)	0.87 (0.03)	0.89	0.90
Vienne KGE	0.74 (0.05)	0.79 (0.11)	0.79	0.83
Average KGE	**0.79** (0.08)	**0.83** (0.08)	**0.85** (0.04)	**0.86** (0.05)
Across Basins SPAEF	**-0.10** (0.45)	**0.41** (0.19)	**0.25** (0.23)	**0.49** (0.15)

The single basin holdout evaluation based on the KGE5 and KSP5 calibrations (Table 4) shows that discharge performances (average of 0.85 and 0.86) are better than the five-basin holdout (KGE1 and KSP1) and very similar to the KGE6 and KSP6 calibrations. Again, the multi-objective calibrations seem more robust for parameter transfer when evaluated against discharge only. For the SPAEF performance evaluation KGE5 performs better than KGE1, indicating better parameter transfer when calibrated against more and diverse basins. However, spatial pattern performances are still considerably better for the ungauged assessment based on multiple objectives in KSP5. Also, KSP5 (SPAEF around 0.5) performs better than KSP1 (SPAEF around 0.4).

In summary, the four ungauged basin tests indicate that discharge can be predicted with average KGEs around 0.79 to 0.83 across the six selected basins based on parameter transfer from calibration of neighboring basins, even when only a single basin is used to estimate parameters for five neighboring basins. . Performances on discharge improve further when including an additional objective function in the form of AET patterns and when calibrating across five basins and evaluating on a single holdout basin. Similarly, spatial patterns can be simulated with average SPAEF values of 0.41 and 0.49, i.e., somewhat lower than KSP6 at 0.61, when only accounting for AET patterns from neighboring basins in the parameter estimation. On the contrary, spatial patterns are very poorly represented when parameters are based on single-basin and single-objective calibrations (KGE1).

In addition to the jack-knifing validation for ungauged basins, a validation test for internal discharge stations was performed for the KGE5 and KSP5 holdout (ungauged) simulations. This test was intended to analyze the possible added value of spatial pattern calibration on internal discharge stations' performance compared to a pure discharge calibration. Since the spatial pattern calibration will not directly influence the temporal performance of the simulated discharge, the internal validation focuses on the discharge bias (Eq. (2); β term) alone and not the KGE.

Since spatial patterns of AET are only included for the period March-November, they are likely to mainly influence the summer water balance where AET has the most impact. Hence, annual and summer statistics are estimated separately. Figure 6 illustrates the location of 46 internal discharge stations and the difference in absolute bias (%) between the ungauged

563 simulations from the KGE5 and KSP5 holdout experiments. For annual statistics (Figure 6
564 top panel), results are very similar (same average bias) and most stations have differences
565 between plus and minus 10%. For the Meuse basin, significant improvements can be detected
566 in the bias for KSP5, while KGE5 tends to be better for the Elbe basin. For the summer
567 statistics (Figure 6 bottom panel) the KSP5 has a slightly lower average bias with
568 considerable improvements for the Meuse and Vienne. At the same time, differences for the
569 Elbe basin are more polarized with stations that are better for both KGE5 and KSP5. Overall,
570 the analysis did not show a clear improvement in biases when constraining the models with
571 spatial patterns in the holdout test. If analyzing KGE and the α and r terms of KGE (Eq. (2)),
572 the KGE-only calibrations performed best for internal station validation in the holdout test.
573 This is illustrated by Figure B1 in the supplementary information section, which shows
574 results for both KGE, and its three components.

575 The model performances presented in this study should be evaluated in light of the
576 uncertainties associated to them. One aspect of this uncertainty is the sampling uncertainty
577 associated with the KGE metric (Clark et al., 2021). The sampling uncertainty represents the
578 uncertainty related to the time window used for the KGE calculation, since the KGE metric is
579 sensitive to the variance of the evaluation period. This uncertainty can be significant and is
580 important especially when evaluating the applicability of a given model for a particular
581 purpose. Even though it is less important for the comparison of different calibration
582 experiments based on the same evaluation periods, the uncertainties associated to each of the
583 evaluation stations used in the study are given in Tables A1 and A2 in Appendix A. The
584 uncertainties are estimated based on the method described in (Clark et al., 2021) and vary
585 between stations but are largely correlated between calibration experiments.

586

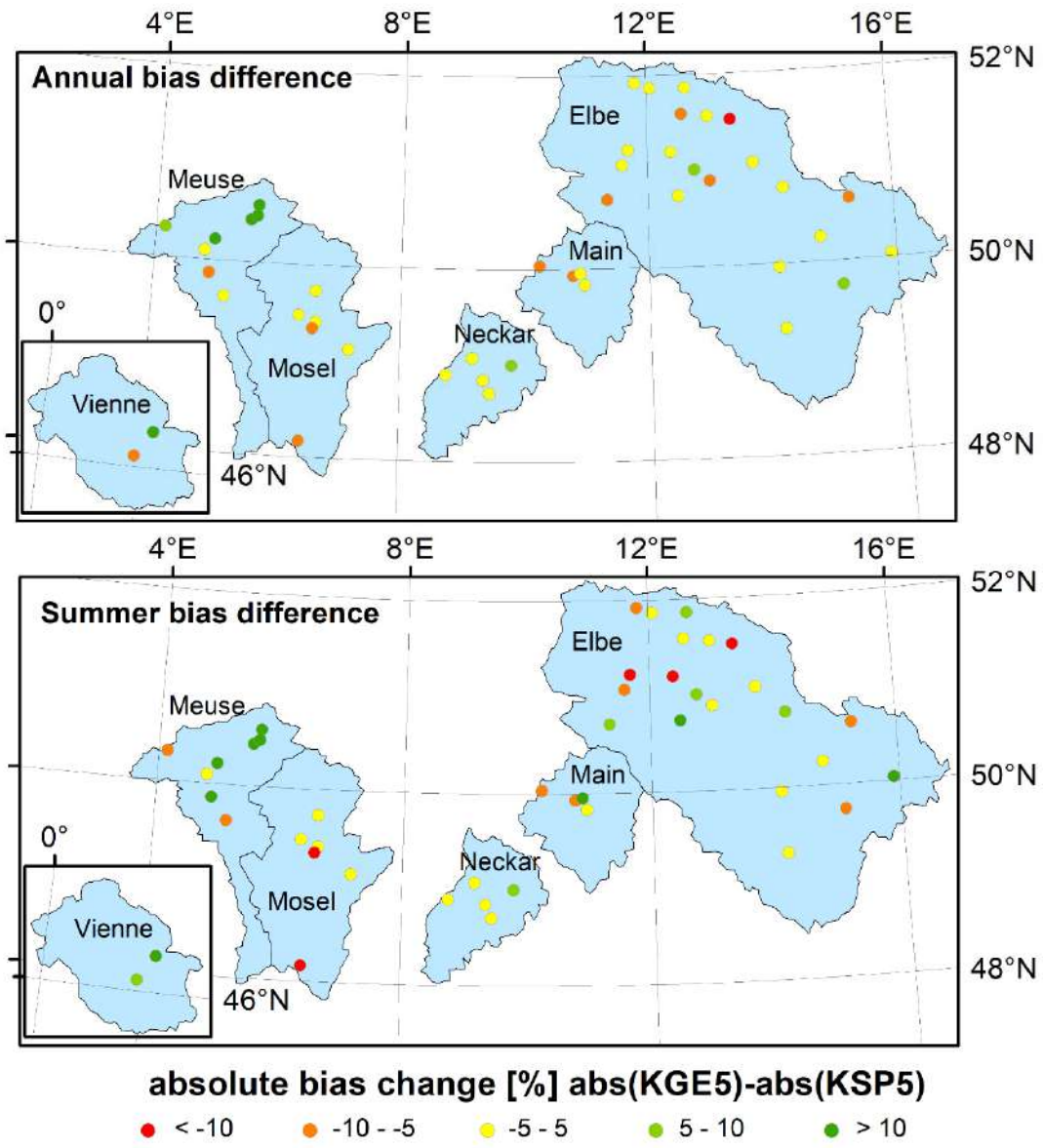


Figure 6: Difference in absolute discharge bias between KGE5 holdout and KSP5 holdout for 46 internal discharge gauges in the six basins for the full year (top) and summer period (May to September) (bottom). Green colors indicate that constraining a model with streamflow and AET leads to better streamflow predictions in ungauged basins than constraining the model with streamflow only. Red colors indicate the opposite.

4 Discussion

The single- versus multi-objective calibration experiment presented here illustrated a minimal tradeoff in discharge performance when adding the spatial pattern-oriented metric to the traditional KGE objective function (**Figure *2***). This result is very similar to previous studies (Demirel, Mai, et al., 2018; Kumar, Samaniego, et al., 2013; Rakovec, Kumar, Mai, et al., 2016; Soltani, Bjerre, et al., 2021; Zink et al., 2018) and can be attributed to two main factors. Firstly, the metric design, with a long-term average bias-insensitive spatial pattern metric introduces limited conflict to matching the discharge biases and no conflict with the temporal dynamics of the discharge simulations. Secondly, single-objective calibrations based on downstream discharge only, are known to constrain the spatial distribution of internal fluxes to a minimal extent (Stisen et al., 2011), causing a high degree of equifinality. Consequently, the addition of a spatial pattern metric can be viewed as a means of selecting the best spatial pattern match among an extensive set of plausible parameter sets (all producing satisfying KGEs). These results on objective function selection, are consistent for both the single-basin and multi-basin tests (with six basins, Figure 3). Not surprisingly, it also becomes evident that a good discharge performance (KGE) does not guarantee a good spatial pattern performance.

In light of the low tradeoff for discharge, single-basin versus multi-basin calibrations, results are best analyzed through comparing the spatial patterns of AET and resulting parameter fields. Here, it becomes clear that single-basin single-objective calibration can select parameter sets that are entirely inconsistent between the basins (Figure 5) and displays internal spatial AET patterns that are reverse of the observed patterns (Figure 4). Interestingly, the multi-basin KGE calibration (KGE6) shows that simply adding multiple basins in this case enables the model to obtain a somewhat realistic spatial pattern without being constrained specifically to AET. However, the spatial metric must be included to improve this pattern and spatial variability (KSP6). Logically, one joint calibration (KGE6 and KSP6) also ensures a spatially consistent parameter field (Figure 5) and thereby also spatially consistent AET patterns (Figure 4). This point has previously been highlighted by Samaniego et al. (2017), who illustrated the shortcomings in producing seamless parameter fields based on multiple single basin calibrations without parameter regionalization across Europe. Eventually, the goal of regional to continental scale distributed hydrologic modelling is to produce scalable spatial patterns of all states and fluxes across the entire model domain.

Moving on to the spatial holdout experiments, first with single basin calibrations (five holdouts) and later with multi-basin calibrations (single holdouts), the parameter transfer to "ungauged" basins results in average KGE values between 0.79 and 0.86 even when transferring parameters from a single basin to five neighboring basins.

For these holdout experiments, the mean KGE for ungauged basins lies around 0.8 (Figure 4) compared to 0.88 for the multi-basin calibrations (KGE6 and KSP6 in Table 3). This is probably a result of a considerable similarity between the basins and their relatively large size, all of them encompassing a range of land use, soil texture, and climate conditions. Also, the six basins were chosen because they all fulfilled the criteria of a similar climate and topography, and previous performance in a Pan-European modeling context (Rakovec, Kumar, Mai, et al., 2016). In this context, the robustness of parameter transferability might be overestimated compared to basins with less similarity.

Other studies have analyzed parameter transferability and KGE performance drop by spatial validation in ungauged basins. A recent and very relevant example is the model intercomparison paper by (Mai et al., 2022). They explicitly performed a spatial validation

test against basins not included in the calibration for a range of different model codes over the Great Lakes region in North America. They reported average loss in KGE of around 0.26 for locally calibrated models using a simple parameter transfer scheme and a loss of 0.10 KGE for regionally calibrated models. In comparison, our study reports a loss of KGE of 0.14 for the KGE1 holdout, 0.07 for KSP1 and 0.03 and 0.02 for the KGE6 and KSP6 holdouts (evaluated through the KGE5 and KSP5 performances).

For the parameter transfer, the experiments including AET during calibration (KSP1 and KSP5) produce better spatial patterns (SPAEF 0.41 and 0.49) when combining ungauged basins, as compared to the KGE-only calibrations (SPAEF -0.10 and 0.25), however KGE5 produced better patterns than KGE1. This is in line with the results of Poméon et al. (2018) who calibrated sparsely gauged basins using remote sensing products. Their study showed that including AET to model calibration significantly improved the performance of the evapotranspiration simulation whereas soil moisture and total water storage predictions were within a good predictive range.

The internal validation against 46 discharge stations was intended to evaluate whether adding spatial patterns to the calibration would improve the discharge bias performance within each basin. Somewhat surprisingly and discouraging, such a systematic bias improvement could not be verified. A previous study by Conradt et al. (2013) on the Elbe basin revealed large discrepancies between water balance AET (precipitation-discharge) and remote sensing-based AET on the sub-basin level. This could indicate that sub-basin water balances are in some cases largely controlled by factors other than AET. This could be water divergence, abstraction, or inter-basin groundwater flow (Le Mesnil et al., 2020; Soltani, Koch, et al., 2021). Wan et al. (2015) showed that the inter-basin transfer of water could cause significant errors in the water balance-based AET calculations. Alternatively, the accuracy of the satellite-based AET might not be sufficient to describe differences at the sub-basin level. Recent analyses, using the AET dataset used in this study, have demonstrated that remote sensing-based AET can reproduce large-scale AET patterns across major European basins (> 25.000 km^2) (Stisen et al., 2021), while studies like Conradt et al. (2013) and Soltani et al. (2021) indicate substantial deviations for smaller sub-basins (below 200 to 500 km^2).

5 Conclusions

The need for systematically transferring parameters to ungauged basins while respecting their landscape heterogeneity and water balance motivated us to expand our previous single-basin experiments (Demirel, Mai, et al., 2018) to a regional scale study. In this study, we elaborated on the value of multi-basin, multi-objective model calibration for distributed hydrologic modelers incorporating readily available global remote sensing data in flexible open-source models with cutting-edge parameter regionalization schemes like the multi-parameter regionalization in mHM.

We first selected the most relevant parameters for spatial calibration using a sensitivity analysis. Then remotely sensed AET based on the two-source energy budget approach is used together with outlet discharge time series to constrain mHM simulations. Through a series of calibration and cross-validation experiments we identify tradeoffs between objective functions and examine the robustness of parameter transferability to ungauged basins.

We can draw the following conclusions from our results:

- Multi-objective calibrations for both individual and multiple basins resulted in balanced solutions leading to better spatio-temporal performances compared to single-objective calibrations. Adding new constraints on spatial patterns only lead to a very limited deterioration in discharge performance while they improve the model predictions for actual evapotranspiration.
- Combining multi-basin and multi-objective calibration has positive impacts on the simulated fluxes and improves the spatial consistency of parameter fields and their transferability to ungauged basins.

Multi-basin calibration is found to be the most crucial element of robust parametrizations if only focusing on discharge. However, adding spatial pattern objectives further ensures spatial consistency, performance, and transferability. Improved model parametrizations in distributed hydrologic models via different transfer functions in combination with appropriate spatial calibration frameworks could facilitate the applications of global hyper-resolution models for “everywhere” (Bierkens et al., 2015) and “without an illogical (unseamless) patchwork of states and fluxes” (Mizukami et al., 2017) in the future. Future work should incorporate more than six basins and spatial patterns of other variables readily available from reliable satellite products.

Appendix A: Results of the the jackknife and bootstrap based sampling uncertainty analysis.

Clark et al (Clark et al., 2021) showed that popular temporal metrics in hydrology, i.e. NSE and KGE, are often subject to inevitable sampling uncertainty. This is due to the fact that differences between observed and simulated streamflow values at random time steps in time series can have significant effects on the overall metric value (Knoben & Spieler, 2022). Therefore, we assessed the sampling uncertainty in KGE results of KGE1, KGE6, KSP1 and KSP6 cases presented in Table 3, using the gumboot R package (Clark et al., 2021) which utilize a jackknife-after-bootstrap method of Efron (1992) to estimate standard errors (SEJaB). Note that this has been done for all 46+6 (validation+calibration) stations listed in Table A1. Uncertainty is represented as confidence interval i.e. the 5th to 95th percentile of the bootstrap samples. Correlation analysis between the SEJaB scores across all 52 stations gives an R^2 of 0.68 for KGE1 vs. KSP1 and 0.76 for KGE6 vs. KSP6. This indicates that the uncertainties are largely related to the specific stations and the variance and error structure of the hydrograph.

Table A1: Sampling uncertainty of KGE metric for KGE1 and KSP1 cases. Rows in bold indicate the six downstream stations used for calibration.

GRDC station	**KGE1**						**KSP1**				
	Basin	**p05**	**p50**	**p95**	**score**	**seJab**	**p05**	**p50**	**p95**	**score**	**seJab**
6340180	**Elbe**	**0.843**	**0.883**	**0.901**	**0.894**	**0.012**	**0.807**	**0.854**	**0.876**	**0.864**	**0.022**
6340130	Elbe	0.711	0.775	0.860	0.780	0.036	0.628	0.697	0.796	0.701	0.037
6340170	Elbe	0.766	0.841	0.914	0.844	0.042	0.700	0.775	0.856	0.777	0.023
6340300	Elbe	0.251	0.389	0.559	0.392	0.135	0.155	0.305	0.487	0.311	0.173
6340190	Elbe	0.699	0.764	0.853	0.768	0.042	0.596	0.669	0.776	0.673	0.039
6340600	Elbe	0.654	0.709	0.799	0.714	0.049	0.619	0.694	0.813	0.699	0.057
6340700	Elbe	0.057	0.357	0.622	0.374	0.219	0.026	0.329	0.557	0.356	0.139
6340200	Elbe	0.027	0.166	0.329	0.169	0.097	-0.122	0.026	0.187	0.032	0.121
6340320	Elbe	0.485	0.630	0.754	0.629	0.092	0.565	0.705	0.819	0.709	0.085
6340365	Elbe	0.245	0.402	0.505	0.405	0.083	-0.144	0.099	0.233	0.101	0.120
6340620	Elbe	0.627	0.698	0.819	0.700	0.047	0.604	0.700	0.863	0.704	0.069
6340120	Elbe	0.694	0.760	0.856	0.766	0.054	0.588	0.659	0.770	0.663	0.053
6340630	Elbe	0.490	0.531	0.596	0.534	0.030	0.372	0.422	0.497	0.423	0.047
6140400	Elbe	0.664	0.732	0.834	0.738	0.067	0.558	0.634	0.750	0.638	0.054
6340621	Elbe	0.618	0.679	0.767	0.679	0.034	0.539	0.641	0.799	0.645	0.053
6140500	Elbe	0.615	0.653	0.693	0.655	0.021	0.505	0.559	0.616	0.560	0.016
6140481	Elbe	0.772	0.823	0.857	0.825	0.020	0.703	0.750	0.792	0.752	0.017
6140600	Elbe	0.343	0.459	0.605	0.458	0.051	0.532	0.638	0.755	0.637	0.058
6140250	Elbe	0.381	0.504	0.663	0.504	0.084	0.266	0.378	0.527	0.375	0.060
6140450	Elbe	0.225	0.351	0.456	0.354	0.056	0.237	0.326	0.407	0.325	0.015
6140300	Elbe	0.595	0.643	0.743	0.646	0.068	0.366	0.436	0.579	0.437	0.087
6340302	Elbe	0.808	0.846	0.875	0.855	0.024	0.719	0.780	0.837	0.784	0.027
6335500	**Main**	**0.904**	**0.929**	**0.944**	**0.939**	**0.010**	**0.889**	**0.913**	**0.930**	**0.921**	**0.007**
6335301	Main	0.905	0.932	0.945	0.941	0.016	0.900	0.924	0.938	0.932	0.015
6335303	Main	0.902	0.926	0.941	0.931	0.010	0.893	0.922	0.937	0.925	0.025
6335530	Main	0.678	0.743	0.802	0.739	0.035	0.666	0.736	0.779	0.732	0.055

6335800	Main	0.719	0.756	0.794	0.762	0.018	0.704	0.746	0.789	0.751	0.019
6421101	**Meuse**	**0.923**	**0.946**	**0.958**	**0.956**	**0.008**	**0.899**	**0.924**	**0.936**	**0.932**	**0.008**
6221500	Meuse	0.829	0.861	0.896	0.872	0.020	0.785	0.826	0.856	0.835	0.022
6221680	Meuse	0.835	0.890	0.929	0.895	0.029	0.653	0.715	0.780	0.720	0.040
6221102	Meuse	0.689	0.748	0.798	0.754	0.034	0.648	0.720	0.772	0.724	0.049
6121240	Meuse	0.056	0.284	0.488	0.273	0.052	-0.120	0.150	0.412	0.134	0.099
6221550	Meuse	0.804	0.827	0.841	0.838	0.016	0.791	0.846	0.881	0.858	0.029
6221120	Meuse	0.782	0.862	0.897	0.874	0.046	0.716	0.804	0.845	0.819	0.039
6221620	Meuse	0.341	0.421	0.488	0.420	0.050	0.340	0.446	0.536	0.445	0.064
6221200	Meuse	0.744	0.828	0.896	0.832	0.037	0.630	0.719	0.799	0.721	0.042
6336050	**Mosel**	**0.921**	**0.951**	**0.964**	**0.960**	**0.008**	**0.892**	**0.932**	**0.949**	**0.943**	**0.025**
6336500	Mosel	0.885	0.930	0.954	0.935	0.027	0.872	0.911	0.933	0.920	0.026
6336800	Mosel	0.690	0.769	0.865	0.774	0.055	0.721	0.807	0.882	0.816	0.037
6336900	Mosel	0.782	0.838	0.882	0.832	0.036	0.780	0.851	0.909	0.845	0.035
6336920	Mosel	0.042	0.207	0.370	0.209	0.051	0.252	0.407	0.541	0.403	0.096
6336910	Mosel	0.743	0.791	0.840	0.793	0.020	0.719	0.794	0.848	0.786	0.048
6136200	Mosel	0.405	0.548	0.703	0.557	0.062	0.214	0.385	0.570	0.395	0.063
6335600	**Neckar**	**0.891**	**0.931**	**0.948**	**0.942**	**0.016**	**0.872**	**0.911**	**0.926**	**0.921**	**0.014**
6335601	Neckar	0.863	0.911	0.927	0.919	0.022	0.841	0.887	0.903	0.896	0.017
6335602	Neckar	0.689	0.752	0.830	0.756	0.034	0.640	0.710	0.796	0.714	0.042
6335660	Neckar	0.733	0.819	0.860	0.825	0.042	0.745	0.802	0.830	0.807	0.045
6335291	Neckar	0.699	0.754	0.792	0.756	0.027	0.774	0.810	0.834	0.814	0.020
6335690	Neckar	0.454	0.512	0.580	0.517	0.042	0.462	0.528	0.606	0.533	0.050
6123400	**Vienne**	**0.832**	**0.892**	**0.916**	**0.913**	**0.018**	**0.814**	**0.882**	**0.906**	**0.899**	**0.025**
6123450	Vienne	0.072	0.279	0.457	0.286	0.080	0.328	0.528	0.663	0.531	0.100
6123820	Vienne	0.617	0.803	0.863	0.802	0.151	0.684	0.806	0.845	0.825	0.091

724

725

726 Table A2: Sampling Uncertainty of KGE Metric for KGE6 and KSP6 cases. Stations in bold
727 are the six downstream stations used for calibration.

728

GRDC station		KGE6					KSP6				
	Basin	p05	p50	p95	score	seJab	p05	p50	p95	score	seJab
6340180	**Elbe**	**0.824**	**0.857**	**0.877**	**0.865**	**0.011**	**0.782**	**0.820**	**0.844**	**0.828**	**0.021**
6340130	Elbe	0.688	0.769	0.861	0.770	0.045	0.694	0.758	0.848	0.761	0.043
6340170	Elbe	0.774	0.834	0.903	0.838	0.021	0.749	0.818	0.891	0.821	0.029
6340300	Elbe	0.197	0.332	0.473	0.332	0.118	0.021	0.185	0.351	0.188	0.149
6340190	Elbe	0.676	0.762	0.868	0.763	0.049	0.678	0.748	0.854	0.751	0.049
6340600	Elbe	0.698	0.754	0.841	0.758	0.043	0.667	0.723	0.813	0.728	0.047
6340700	Elbe	0.118	0.428	0.682	0.435	0.262	-0.065	0.245	0.531	0.255	0.187
6340200	Elbe	-0.020	0.110	0.236	0.113	0.103	-0.266	-0.092	0.072	-0.089	0.123
6340320	Elbe	0.434	0.578	0.700	0.577	0.076	0.429	0.574	0.700	0.578	0.091
6340365	Elbe	0.101	0.263	0.356	0.268	0.101	-0.361	-0.030	0.138	-0.031	0.146
6340620	Elbe	0.658	0.734	0.833	0.735	0.035	0.637	0.715	0.839	0.717	0.063
6340120	Elbe	0.675	0.759	0.870	0.761	0.059	0.670	0.738	0.849	0.743	0.066
6340630	Elbe	0.517	0.558	0.621	0.560	0.039	0.413	0.455	0.519	0.457	0.052
6140400	Elbe	0.642	0.732	0.851	0.733	0.059	0.642	0.716	0.833	0.720	0.071
6340621	Elbe	0.662	0.731	0.822	0.732	0.030	0.571	0.660	0.813	0.664	0.063
6140500	Elbe	0.649	0.688	0.729	0.691	0.021	0.540	0.588	0.642	0.590	0.022
6140481	Elbe	0.822	0.862	0.883	0.864	0.020	0.761	0.805	0.843	0.808	0.014
6140600	Elbe	0.514	0.610	0.732	0.608	0.048	0.607	0.702	0.785	0.706	0.054
6140250	Elbe	0.309	0.459	0.640	0.458	0.080	0.357	0.466	0.621	0.466	0.068
6140450	Elbe	0.054	0.157	0.246	0.154	0.025	-0.001	0.097	0.189	0.095	0.024
6140300	Elbe	0.540	0.616	0.767	0.617	0.078	0.446	0.510	0.630	0.512	0.068
6340302	Elbe	0.783	0.833	0.866	0.841	0.024	0.769	0.813	0.852	0.823	0.026
6335500	**Main**	**0.807**	**0.842**	**0.877**	**0.843**	**0.019**	**0.824**	**0.860**	**0.893**	**0.862**	**0.021**
6335301	Main	0.797	0.838	0.882	0.839	0.018	0.820	0.860	0.896	0.861	0.020
6335303	Main	0.777	0.818	0.858	0.816	0.023	0.806	0.845	0.884	0.843	0.022
6335530	Main	0.525	0.591	0.657	0.589	0.033	0.561	0.633	0.697	0.629	0.033
6335800	Main	0.832	0.867	0.903	0.873	0.019	0.781	0.835	0.883	0.838	0.019
6421101	**Meuse**	**0.871**	**0.908**	**0.939**	**0.911**	**0.017**	**0.874**	**0.914**	**0.943**	**0.918**	**0.021**
6221500	Meuse	0.777	0.817	0.850	0.823	0.021	0.824	0.853	0.875	0.862	0.012
6221680	Meuse	0.764	0.826	0.884	0.831	0.033	0.891	0.923	0.936	0.933	0.010
6221102	Meuse	0.748	0.809	0.868	0.815	0.028	0.749	0.805	0.864	0.814	0.028
6121240	Meuse	0.174	0.386	0.581	0.372	0.049	0.123	0.335	0.534	0.323	0.054
6221550	Meuse	0.770	0.795	0.812	0.803	0.010	0.786	0.817	0.839	0.829	0.016
6221120	Meuse	0.737	0.812	0.857	0.824	0.042	0.762	0.808	0.840	0.818	0.029
6221620	Meuse	0.397	0.473	0.535	0.471	0.045	0.341	0.414	0.474	0.412	0.043
6221200	Meuse	0.690	0.774	0.852	0.778	0.036	0.784	0.850	0.895	0.855	0.037
6336050	**Mosel**	**0.845**	**0.894**	**0.921**	**0.897**	**0.029**	**0.837**	**0.891**	**0.923**	**0.895**	**0.027**
6336500	Mosel	0.833	0.893	0.934	0.896	0.026	0.828	0.893	0.931	0.896	0.023
6336800	Mosel	0.672	0.764	0.870	0.771	0.059	0.674	0.769	0.872	0.777	0.057
6336900	Mosel	0.820	0.860	0.891	0.859	0.027	0.796	0.845	0.879	0.843	0.036
6336920	Mosel	0.422	0.527	0.623	0.528	0.032	0.453	0.565	0.662	0.566	0.046
6336910	Mosel	0.702	0.741	0.778	0.741	0.012	0.627	0.690	0.741	0.684	0.028

6136200	Mosel	0.268	0.418	0.585	0.427	0.071	0.156	0.314	0.493	0.322	0.068
6335600	**Neckar**	**0.863**	**0.893**	**0.907**	**0.902**	**0.011**	**0.851**	**0.885**	**0.904**	**0.895**	**0.012**
6335601	Neckar	0.831	0.868	0.883	0.877	0.016	0.816	0.859	0.880	0.868	0.017
6335602	Neckar	0.652	0.727	0.817	0.731	0.042	0.616	0.694	0.794	0.699	0.043
6335660	Neckar	0.767	0.852	0.898	0.848	0.082	0.795	0.834	0.867	0.844	0.013
6335291	Neckar	0.702	0.735	0.759	0.734	0.016	0.747	0.788	0.812	0.790	0.017
6335690	Neckar	0.445	0.503	0.581	0.510	0.057	0.433	0.495	0.576	0.501	0.054
6123400	**Vienne**	**0.750**	**0.848**	**0.922**	**0.852**	**0.036**	**0.747**	**0.850**	**0.904**	**0.853**	**0.049**
6123450	Vienne	0.475	0.623	0.703	0.631	0.092	0.547	0.660	0.708	0.679	0.087
6123820	Vienne	0.676	0.801	0.844	0.810	0.061	0.616	0.764	0.838	0.766	0.103

Appendix B: Validation against 46 internal discharge stations

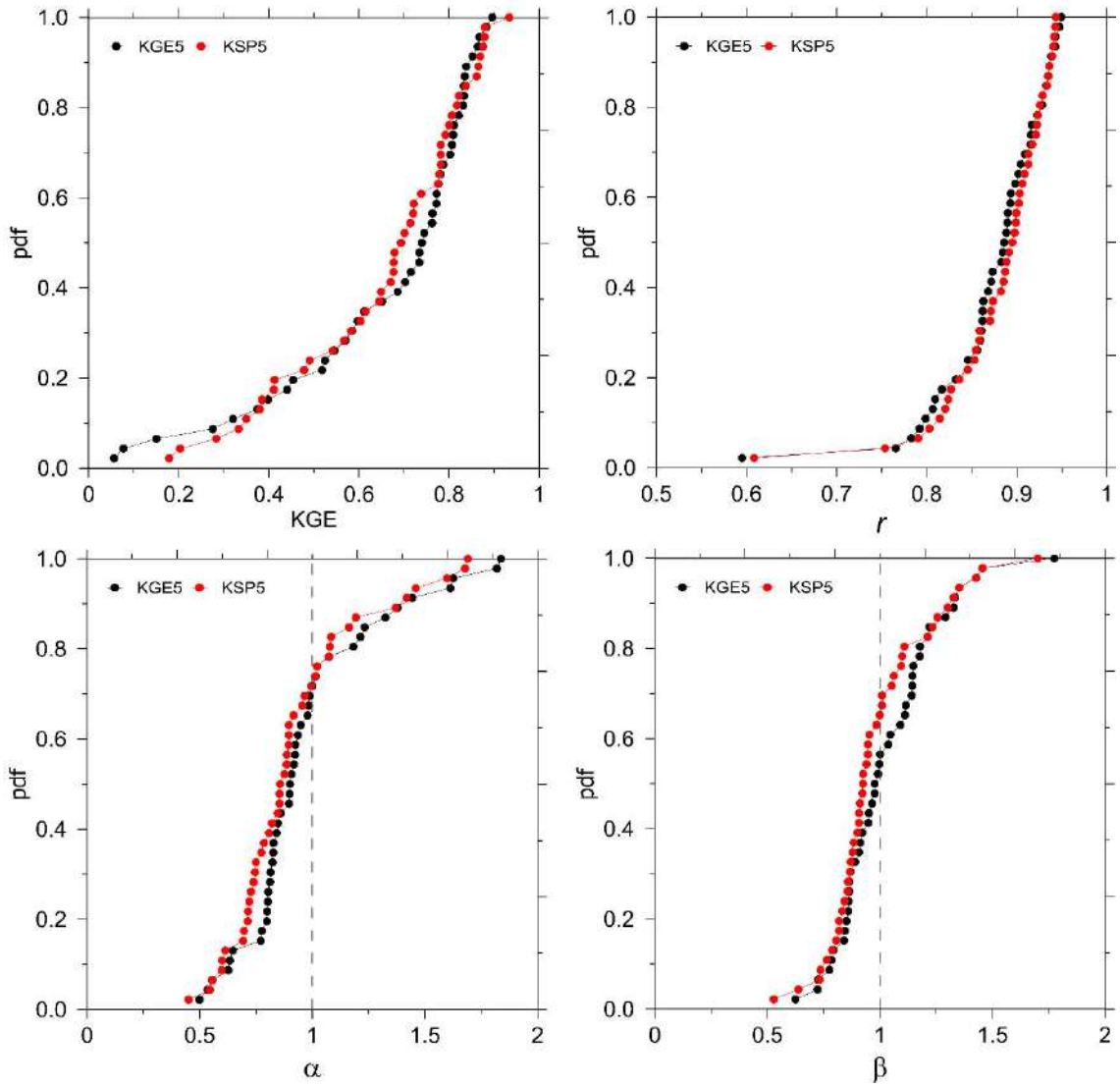


Figure B1: Ranked scores for 46 internal validation stations for KGE5 and KSP5 spatial holdouts. Values are the KGE and its three components *r,* α and β. Data for β correspond to results mapped in Figure 6 (top).

Acknowledgements, Samples, and Data

Data Availability Statement: Discharge data is provided by GRDC data portal (https://portal.grdc.bafg.de/) in Koblenz, Germany. MODIS MOD16A2 v061 product was retrieved from https://doi.org/10.5067/MODIS/MOD16A2.061. SRTM DEM data was retrieved from https://www.earthdata.nasa.gov. The source code of the mHM is publicly available at https://doi.org/10.5281/zenodo.4575390. The source code of the SPAEF metric is publicly available at https://doi.org/10.5281/zenodo.5861253. The source code to quantify the sampling uncertainty in performance metrics (the "gumboot" package) is available at https://github.com/CH-Earth/gumboot. The model calibration software Ostrich is available from https://github.com/usbr/ostrich or http://www.civil.uwaterloo.ca/envmodelling/Ostrich.html. The data for the mHM model simulations are publicly available at https://doi.org/10.5281/zenodo.7931276.

Acknowledgements: We acknowledge the financial support for the SPACE project by the Villum Foundation (http://villumfonden.dk/) through their Young Investigator Program (grant VKR023443). The first author is supported by the Scientific Research Projects Department of Istanbul Technical University (ITU-BAP) under grant number MDA-2022-43762 and the National Center for High Performance Computing of Turkey (UHeM) under grant number 1007292019.

Conflicts of Interest: "The authors declare no conflict of interest."

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

References

Asadzadeh, M., & Tolson, B. (2013). Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. *Engineering Optimization*, *45*(12), 1489–1509. https://doi.org/10.1080/0305215X.2012.748046

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Beume, N., & Rudolph, G. (2006). Faster S-metric calculation by considering dominated hypervolume as Klee's measure problem. *Proceedings of the 2nd IASTED International Conference on Computational Intelligence, CI 2006*, 231–236. https://doi.org/10.17877/DE290R-12786

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., et al. (2015). Hyper-resolution global hydrological modelling: what is next? *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Busari, I. O., Demirel, M. C., & Newton, A. (2021). Effect of Using Multi-Year Land Use Land Cover and Monthly LAI Inputs on the Calibration of a Distributed Hydrologic Model. *Water*, *13*(11), 1538. https://doi.org/10.3390/w13111538

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The Abuse of Popular Performance Metrics in Hydrologic Modeling. *Water Resources Research*, *57*(9). https://doi.org/10.1029/2020WR029001

Conradt, T., Wechsung, F., & Bronstert, A. (2013). Three perceptions of the evapotranspiration landscape: comparing spatial patterns from a distributed hydrological model, remotely sensed surface temperatures, and sub-basin water balances. *Hydrology and Earth System Sciences*, *17*(7), 2947–2966. https://doi.org/10.5194/hess-17-2947-2013

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017JD028200

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR026085

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018

Demirel, M. C., Koch, J., Mendiguren, G., & Stisen, S. (2018). Spatial Pattern Oriented Multicriteria Sensitivity Analysis of a Distributed Hydrologic Model. *Water*, *10*(9), 1188. https://doi.org/10.3390/w10091188

Doherty, J. (2010). *PEST: Model-Independent Parameter Estimation, User Manual: 5th Edition*. *PEST Manual*.

Efron, B. (1992). Jackknife-After-Bootstrap Standard Errors and Influence Functions.

Journal of the Royal Statistical Society: Series B (Methodological), *54*(1), 83–111. https://doi.org/10.1111/j.2517-6161.1992.tb01866.x

Ekmekcioğlu, Ö., Demirel, M. C., & Booij, M. J. (2022). Effect of data length, spin-up period and spatial model resolution on fully distributed hydrological model calibration in the Moselle basin. *Hydrological Sciences Journal*, *67*(5), 759–772. https://doi.org/10.1080/02626667.2022.2046754

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics*, *45*(2), RG2004. https://doi.org/10.1029/2005RG000183

Feigl, M., Herrnegger, M., Klotz, D., & Schulz, K. (2020). Function Space Optimization: A Symbolic Regression Method for Estimating Parameter Transfer Functions for Hydrological Models. *Water Resources Research*, *56*(10). https://doi.org/10.1029/2020WR027385

Feigl, M., Thober, S., Schweppe, R., Herrnegger, M., Samaniego, L., & Schulz, K. (2022). Automatic Regionalization of Model Parameters for Hydrological Models. *Water Resources Research*, *58*(12). https://doi.org/10.1029/2022WR031966

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., & Freer, J. (2014). Catchment properties, function, and conceptual model representation: Is there a correspondence? *Hydrological Processes*. https://doi.org/10.1002/hyp.9726

George H. Hargreaves, & Zohrab A. Samani. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99. https://doi.org/10.13031/2013.26773

GRDC. (2020). Major River Basins of the World / Global Runoff Data Centre, GRDC. 2nd, rev. ext. ed. Koblenz, Germany: Federal Institute of Hydrology (BfG). Retrieved from https://portal.grdc.bafg.de/

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

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

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

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Knoben, W. J. M., & Spieler, D. (2022). Teaching hydrological modelling: illustrating model structure uncertainty with a ready-to-use computational exercise. *Hydrology and Earth System Sciences*, *26*(12), 3299–3314. https://doi.org/10.5194/hess-26-3299-2022

Ko, A., Mascaro, G., & Vivoni, E. R. (2019). Strategies to Improve and Evaluate Physics-

Based Hyperresolution Hydrologic Simulations at Regional Basin Scales. *Water Resources Research*, *55*(2), 1129–1152. https://doi.org/10.1029/2018WR023521

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873–1886. https://doi.org/10.5194/gmd-11-1873-2018

Koch, J., Demirel, M. C., & Stisen, S. (2022). Climate Normalized Spatial Patterns of Evapotranspiration Enhance the Calibration of a Hydrological Model. *Remote Sensing*, *14*(2), 315. https://doi.org/10.3390/rs14020315

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*(1), 360–379. https://doi.org/10.1029/2012WR012195

Kumar, R., Livneh, B., & Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, *49*(9), 5700–5714. https://doi.org/10.1002/wrcr.20431

Lane, R. A., Freer, J. E., Coxon, G., & Wagener, T. (2021). Incorporating Uncertainty Into Multiscale Parameter Regionalization to Evaluate the Performance of Nationally Consistent Parameter Fields for a Hydrological Model. *Water Resources Research*, *57*(10). https://doi.org/10.1029/2020WR028393

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., et al. (2022). The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL). *Hydrology and Earth System Sciences*, *26*(13), 3537–3572. https://doi.org/10.5194/hess-26-3537-2022

Martinsen, G., He, X., Koch, J., Guo, W., Refsgaard, J. C., & Stisen, S. (2022). Large-scale hydrological modeling in a multi-objective uncertainty framework – Assessing the potential for managed aquifer recharge in the North China Plain. *Journal of Hydrology: Regional Studies*, *41*, 101097. https://doi.org/10.1016/j.ejrh.2022.101097

Matott, L. S. (2017). OSTRICH: an Optimization Software Tool, Documentation and User's Guide. *University at Buffalo Center for Computational Research*, *Version 17*, 79.

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., et al. (2023). Can Hydrological Models Benefit From Using Global Soil Moisture, Evapotranspiration, and Runoff Products as Calibration Targets? *Water Resources Research*, *59*(2). https://doi.org/10.1029/2022WR032064

Le Mesnil, M., Charlier, J.-B., Moussa, R., Caballero, Y., & Dörfliger, N. (2020). Interbasin groundwater flow: Characterization, role of karst areas, impact on annual water balance and flood processes. *Journal of Hydrology*, *585*, 124583. https://doi.org/10.1016/j.jhydrol.2020.124583

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining Conceptual Hydrological Models With Multiple Information Sources. *Water Resources Research*, *54*(10), 8332–8362.

https://doi.org/10.1029/2017WR021895

Norman, J. M., Kustas, W. P., & Humes, K. S. (1995). Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology*, *77*(3–4), 263–293. https://doi.org/10.1016/0168-1923(95)02265-Y

Odusanya, A. E., Schulz, K., & Mehdi-Schulz, B. (2022). Using a regionalisation approach to evaluate streamflow simulated by an ecohydrological model calibrated with global land surface evaporation from remote sensing. *Journal of Hydrology: Regional Studies*, *40*, 101042. https://doi.org/10.1016/j.ejrh.2022.101042

Poméon, T., Diekkrüger, B., & Kumar, R. (2018). Computationally Efficient Multivariate Calibration and Validation of a Grid-Based Hydrologic Model in Sparsely Gauged West African River Basins. *Water*, *10*(10), 1418. https://doi.org/10.3390/w10101418

Rakovec, O., Kumar, R., Attinger, S., & Samaniego, L. (2016). Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resources Research*, *52*(10), 7779–7792. https://doi.org/10.1002/2016WR019430

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., et al. (2019). Diagnostic Evaluation of Large-Domain Hydrologic Models Calibrated Across the Contiguous United States. *Journal of Geophysical Research: Atmospheres*, 2019JD030767. https://doi.org/10.1029/2019JD030767

Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*(12), 8418–8431. https://doi.org/10.1002/2012WR013442

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., & Bhatti, H. A. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *Journal of Hydrology*, *505*, 276–290. https://doi.org/10.1016/j.jhydrol.2013.10.006

Running, S. W., Mu, Q., & Zhao, M. (2021). MODIS/terra net evapotranspiration 8-day L4 global 500m SIN grid V061. *NASA EOSDIS Land Process. DAAC*, *5067*. https://doi.org/10.5067/MODIS/MOD16A2.061

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5), W05523. https://doi.org/10.1029/2008WR007327

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., et al. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, *21*(9), 4323–4346. https://doi.org/10.5194/hess-21-4323-2017

Samaniego, L., Brenner, J., Craven, J., Cuntz, M., Dalmasso, G., Demirel, M. C., et al. (2021, February 3). mesoscale Hydrologic Model - mHM v5.11.1. Leipzig. https://doi.org/10.5281/ZENODO.4462822

Schweppe, R., Thober, S., Müller, S., Kelbling, M., Kumar, R., Attinger, S., & Samaniego, L.

(2022). MPR 1.0: a stand-alone multiscale parameter regionalization tool for improved parameter estimation of land surface models. *Geoscientific Model Development*, *15*(2), 859–882. https://doi.org/10.5194/gmd-15-859-2022

Sirisena, T. A. J. G., Maskey, S., & Ranasinghe, R. (2020). Hydrological Model Calibration with Streamflow and Remote Sensing Based Evapotranspiration Data in a Data Poor Basin. *Remote Sensing*, *12*(22), 3768. https://doi.org/10.3390/rs12223768

Soltani, M., Bjerre, E., Koch, J., & Stisen, S. (2021). Integrating remote sensing data in optimization of a national water resources model to improve the spatial pattern performance of evapotranspiration. *Journal of Hydrology*, *603*, 127026. https://doi.org/10.1016/j.jhydrol.2021.127026

Soltani, M., Koch, J., & Stisen, S. (2021). Using a Groundwater Adjusted Water Balance Approach and Copulas to Evaluate Spatial Patterns and Dependence Structures in Remote Sensing Derived Evapotranspiration Products. *Remote Sensing*, *13*(5), 853. https://doi.org/10.3390/rs13050853

Stisen, S., McCabe, M. F., Refsgaard, J. C., Lerer, S., & Butts, M. B. (2011). Model parameter analysis using remotely sensed pattern information in a multi-constraint framework. *Journal of Hydrology*, *409*(1–2), 337–349. https://doi.org/10.1016/j.jhydrol.2011.08.030

Stisen, S., Soltani, M., Mendiguren, G., Langkilde, H., Garcia, M., & Koch, J. (2021). Spatial Patterns in Actual Evapotranspiration Climatologies for Europe. *Remote Sensing*, *13*(12), 2410. https://doi.org/10.3390/rs13122410

Tangdamrongsub, N., Steele-Dunne, S. C., Gunter, B. C., Ditmar, P. G., Sutanudjaja, E. H., Sun, Y., et al. (2017). Improving estimates of water resources in a semi-arid region by assimilating GRACE data into the PCR-GLOBWB hydrological model. *Hydrology and Earth System Sciences*, *21*(4), 2053–2074. https://doi.org/10.5194/hess-21-2053-2017

Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., & Samaniego, L. (2019). The multiscale routing model mRM v1.0: simple river routing at resolutions from 1 to 50 km. *Geoscientific Model Development*, *12*(6), 2501–2521. https://doi.org/10.5194/gmd-12-2501-2019

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Wan, Z., Zhang, K., Xue, X., Hong, Z., Hong, Y., & Gourley, J. J. (2015). Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States. *Water Resources Research*, *51*(8), 6485–6499. https://doi.org/10.1002/2015WR017311

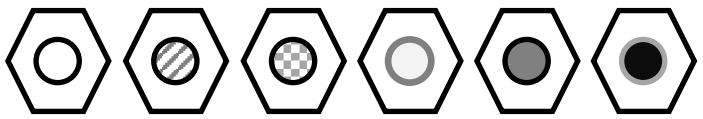
Xiao, M., Mascaro, G., Wang, Z., Whitney, K. M., & Vivoni, E. R. (2022). On the value of satellite remote sensing to reduce uncertainties of regional simulations of the Colorado River. *Hydrology and Earth System Sciences*, *26*(21), 5627–5646. https://doi.org/10.5194/hess-26-5627-2022

Zink, M., Mai, J., Cuntz, M., & Samaniego, L. (2018). Conditioning a Hydrologic Model Using Patterns of Remotely Sensed Land Surface Temperature. *Water Resources Research*, *54*(4), 2976–2998. https://doi.org/10.1002/2017WR021346

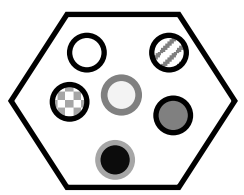
Figure 1.

KGE only

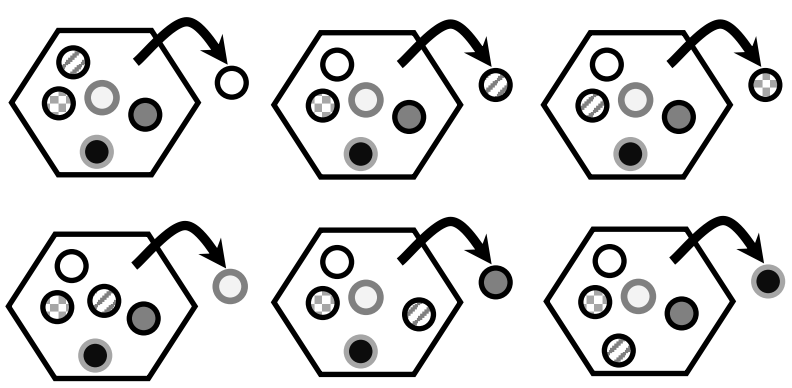
Single basin calibrations - KGE1



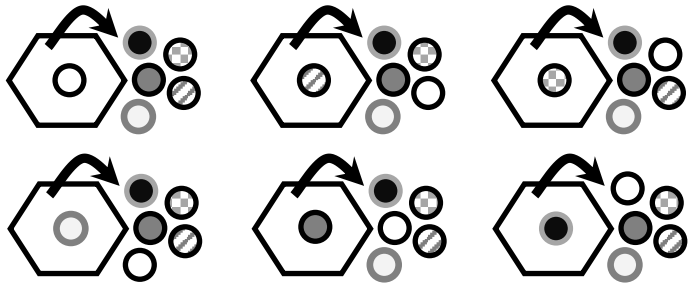
Six basin calibration - KGE6



One basin out calibrations - KGE5

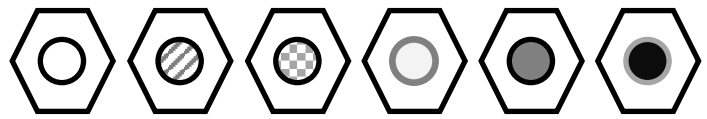


Five-basin out validations - KGE1

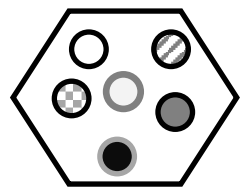


KGE and SPAEF

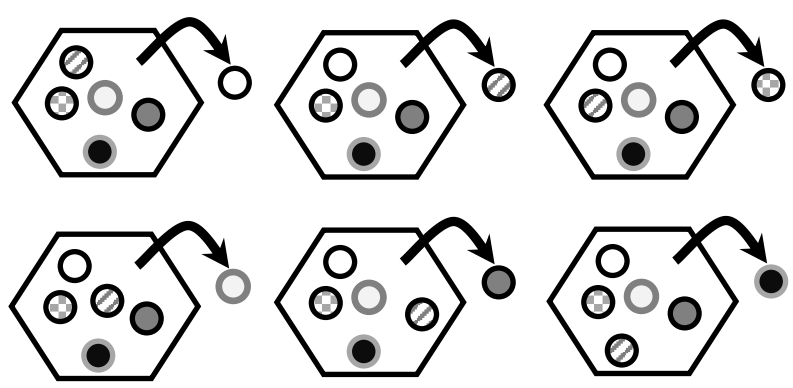
Single basin calibrations - KSP1



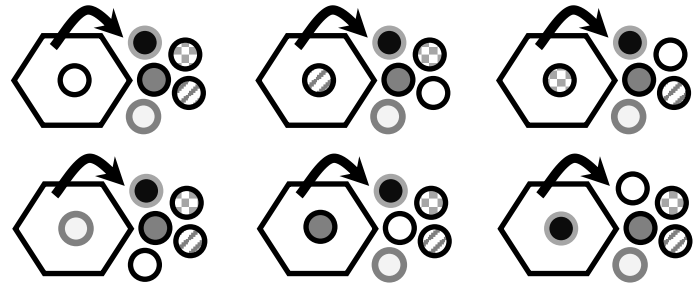
Six basin calibration - KSP6



One basin out calibrations - KSP5



Five-basin out validations - KSP1



Legend

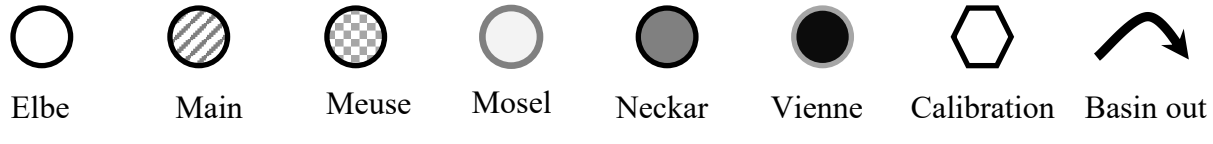


Figure 2.

Elbe

Main

Meuse

Mosel

Neckar

Vienne

Mean SPAEF

$\Sigma(1\text{-KGE})^2$

$\Sigma(1\text{-SPAEF})^2$

KGE

KSP1 (SSR) KGE1 (SSR) KSP1 (Mean) KGE1 (Mean)

Figure 3.

6 Basins

Mean SPAEF

-1 -0.6 -0.2 0.2 0.6 1

0.50 0.40 0.30 0.20 0.10 0.00

$\Sigma(1\text{-KGE})^2$

1 0.9 0.8 0.7 0.6

Mean KGE

0 1 2 3 4 5

$\Sigma(1\text{-SPAEF})^2$

KSP6 (SSR) KGE6 (SSR) KSP6 (Mean) KGE6 (Mean)

Figure 4.

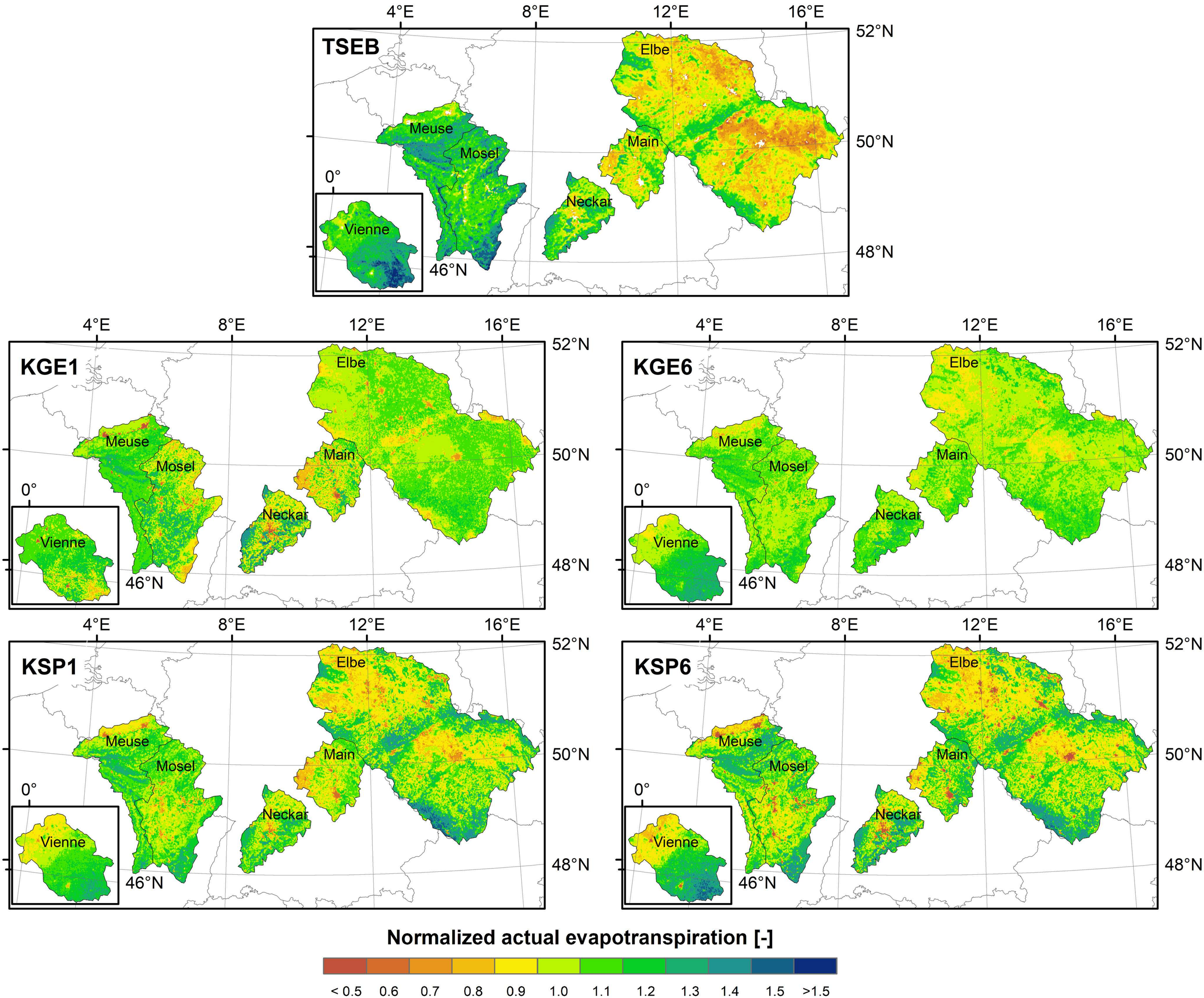

TSEB
KGE1
KGE6
KSP1
KSP6
4°E
8°E
12°E
16°E
52°N
50°N
48°N
46°N
0°
Meuse
Mosel
Elbe
Main
Neckar
Vienne
Normalized actual evapotranspiration [-]
< 0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
>1.5

Figure 5.

KGE1

KSP1

KGE6

KSP6

Elbe
Meuse
Mosel
Main
Neckar
Vienne
0°
4°E
8°E
12°E
16°E
52°N
50°N
48°N
46°N

Normalized field capacity [-]

< 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | > 1.20

KGE1

KSP1

KGE6

KSP6

Elbe
Meuse
Mosel
Main
Neckar
Vienne
0°
4°E
8°E
12°E
16°E
52°N
50°N
48°N
46°N

Crop Coefficient [-]

< 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.10 | 1.20 | 1.40 | > 1.40

Figure 6.

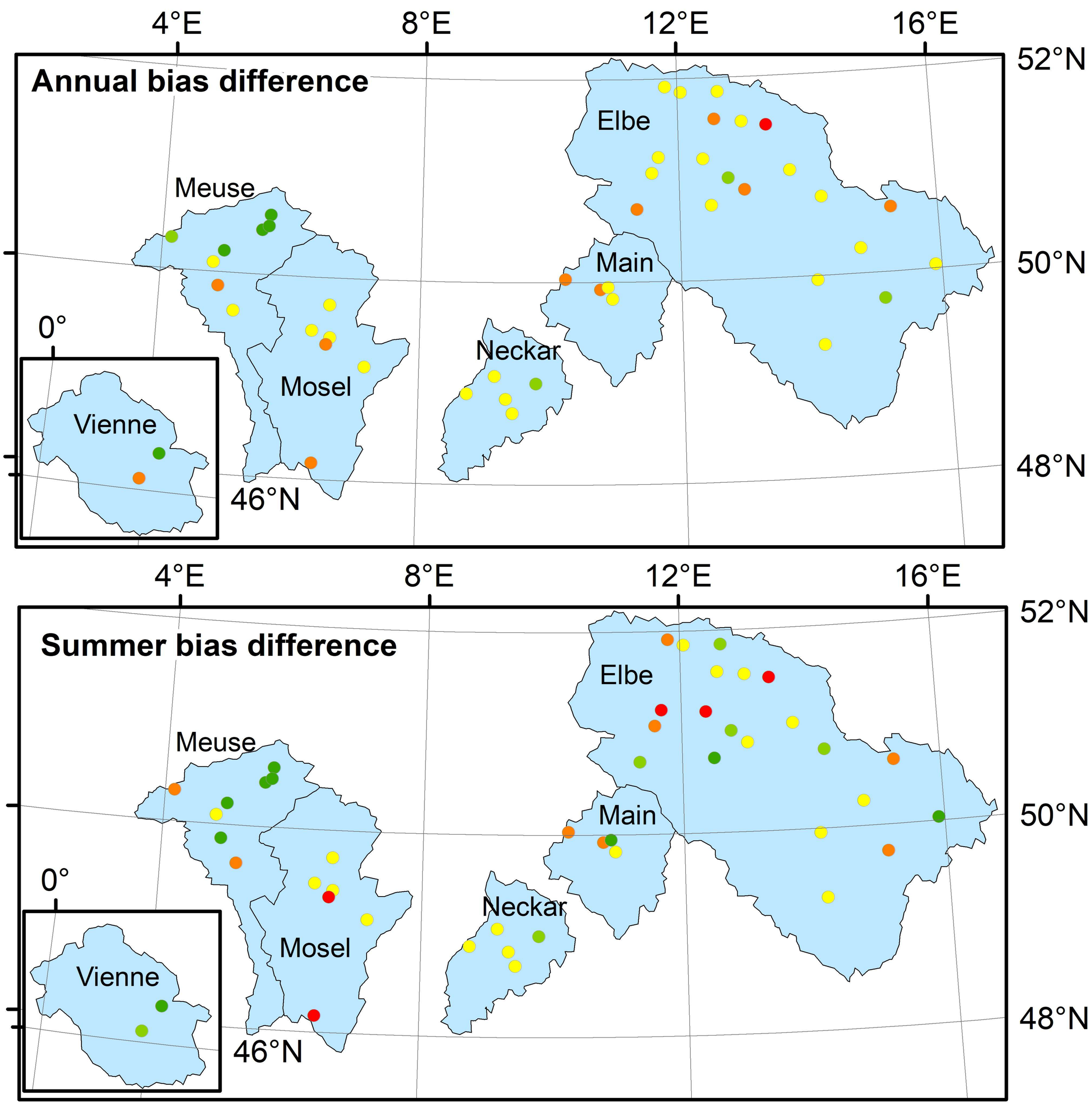
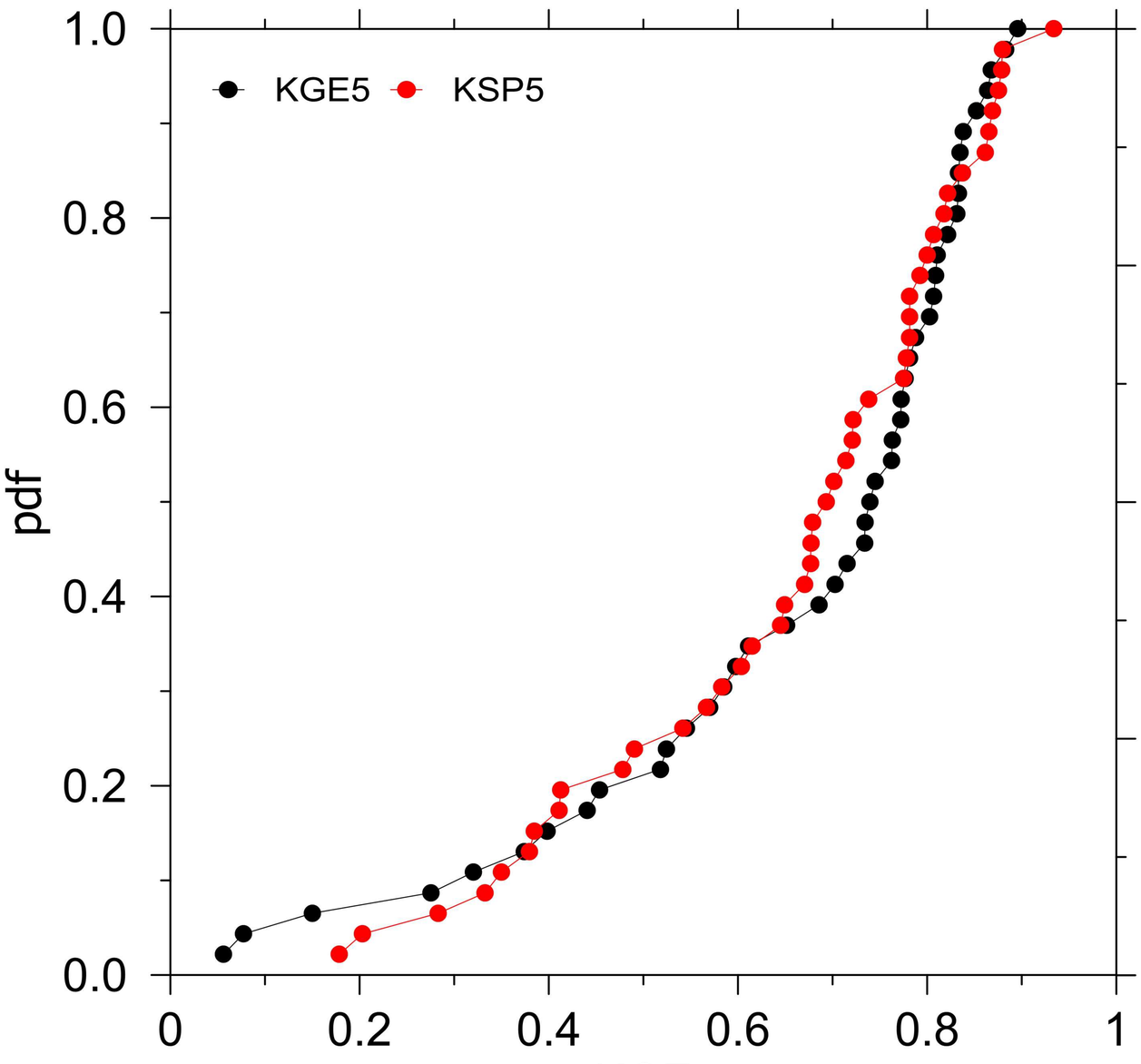
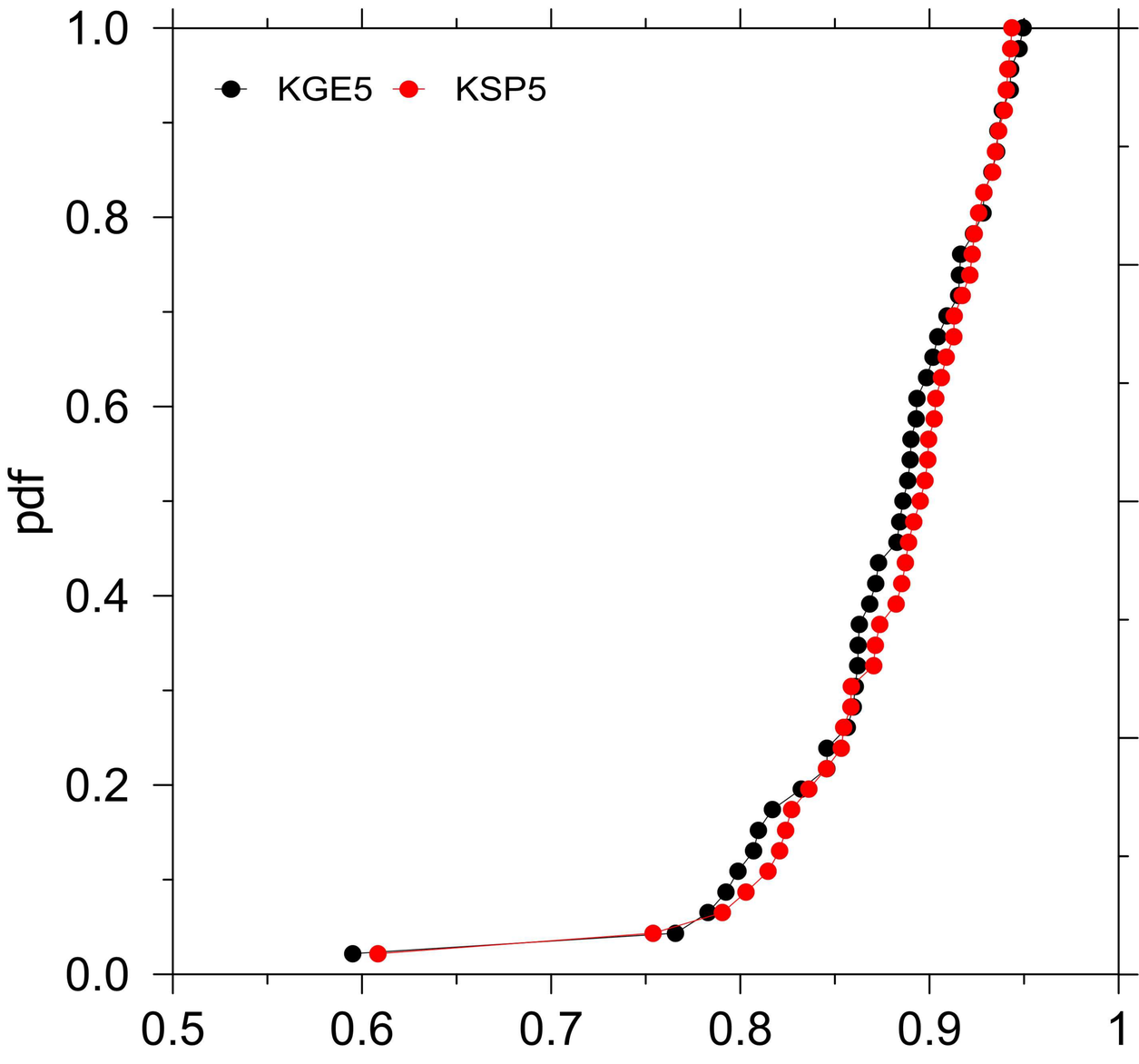
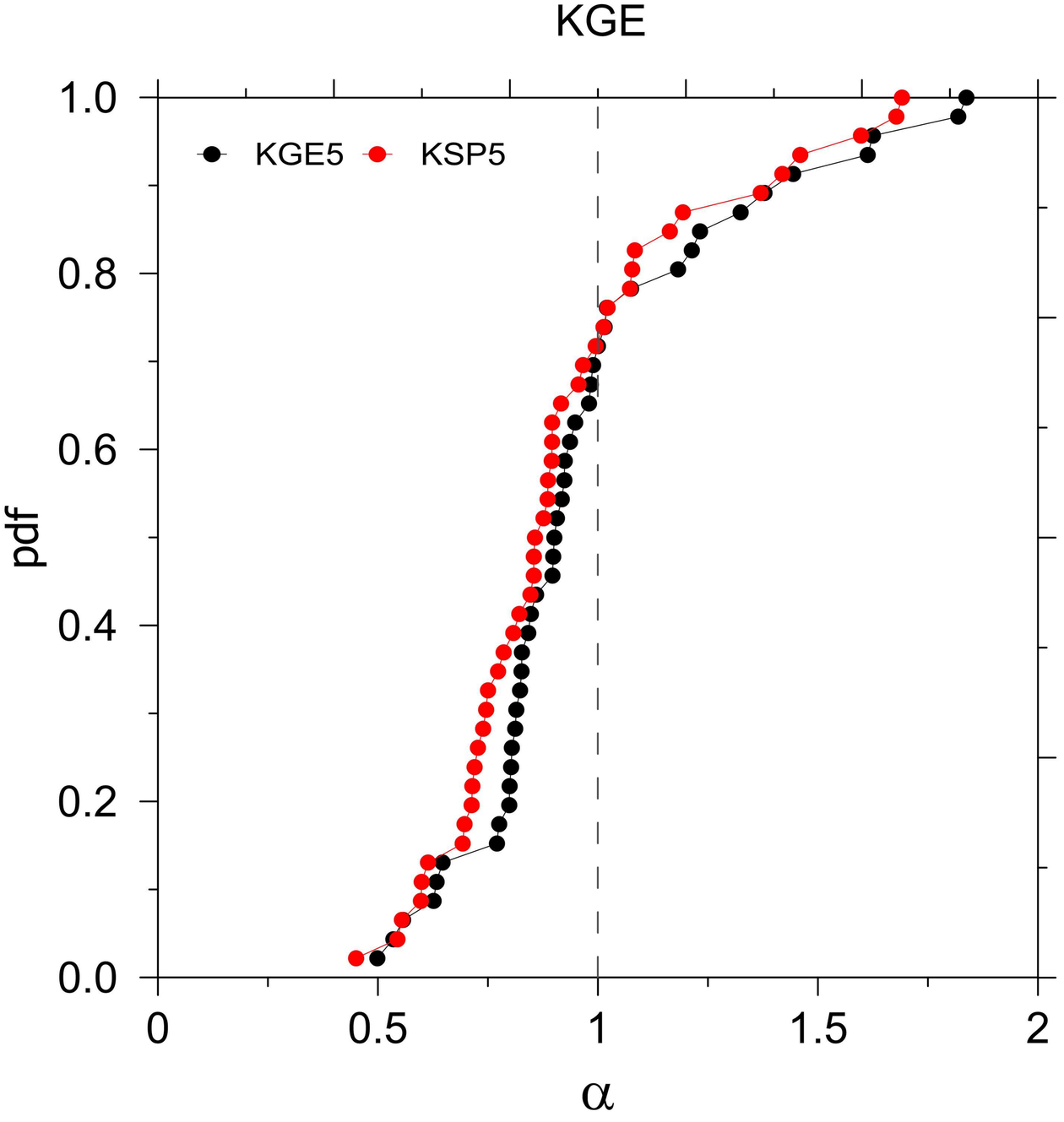
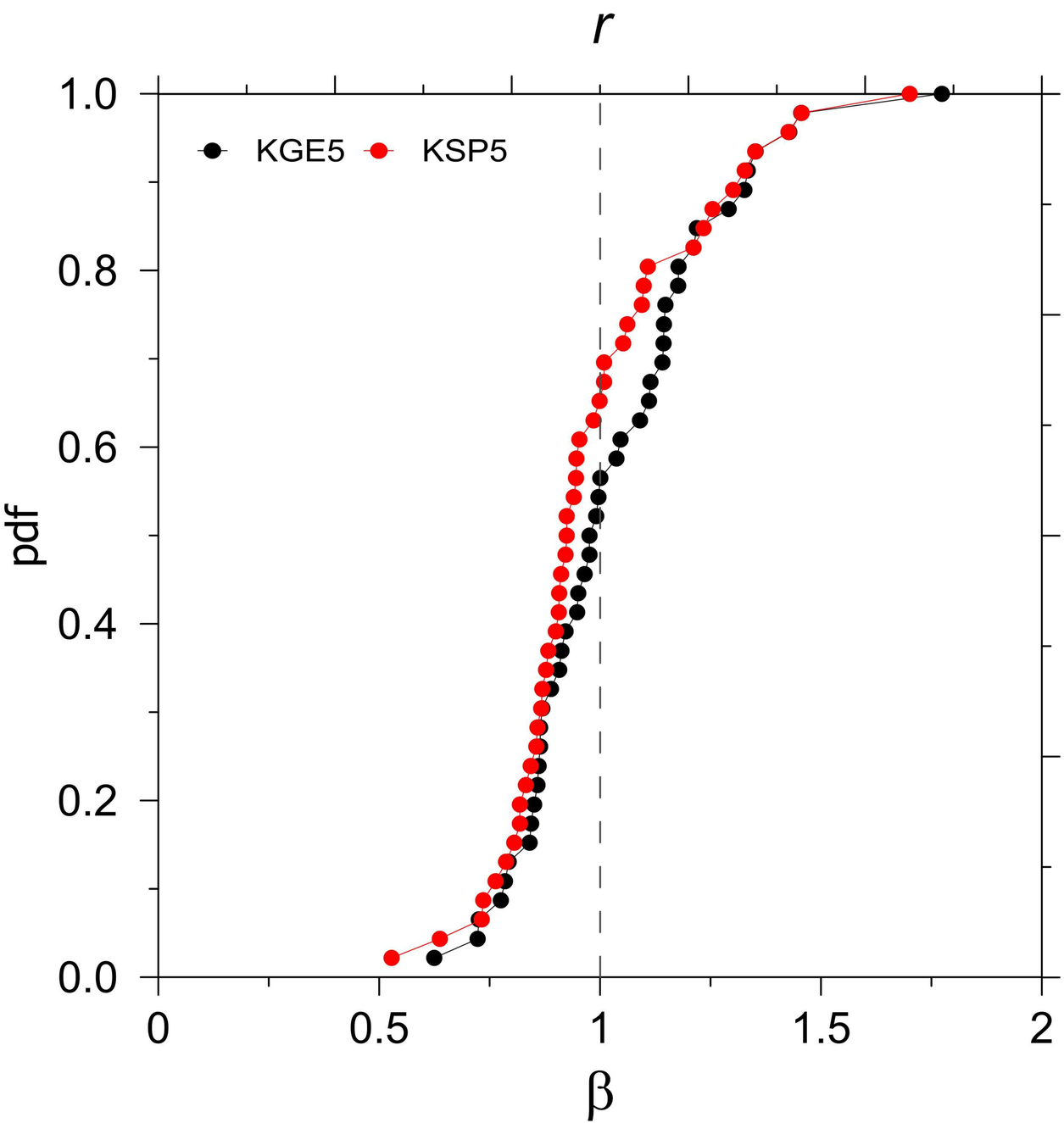

Annual bias difference
4°E
8°E
12°E
16°E
52°N
50°N
48°N
46°N
0°
Meuse
Elbe
Main
Neckar
Mosel
Vienne
Summer bias difference
absolute bias change [%] abs(KGE5)-abs(KSP5)
< -10
-10 - -5
-5 - 5
5 - 10
> 10

Figure B1.


KGE5
KSP5
pdf
1.0
0.8
0.6
0.4
0.2
0.0
0
0.2
0.4
0.6
0.8
1


KGE5
KSP5
pdf
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.6
0.7
0.8
0.9
1


KGE
KGE5
KSP5
pdf
1.0
0.8
0.6
0.4
0.2
0.0
0
0.5
1
1.5
2
α


r
KGE5
KSP5
pdf
1.0
0.8
0.6
0.4
0.2
0.0
0
0.5
1
1.5
2
β