# Benchmarking multi-component spatial metrics for hydrologic model calibration using MODIS AET and LAI products

**E. B. Yorulmaz[1], E. Kartal[1] and M. C. Demirel[1]**

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

Corresponding author: Eymen Berkay Yorulmaz (yorulmaz21@itu.edu.tr, ORCID 0000-0003-3370-9465)

**Key Points:**

- Newly proposed spatial metrics offer significant improvements in discriminating between two raster maps
- Selecting appropriate spatial metric proved to be very crucial even for the global search algorithms
- Sampling uncertainty in metrics increases with newly added components

## Abstract

SPAtial EFficiency (SPAEF) metric is one of the most thoroughly metrics in hydrologic community. In this study, our aim is to improve SPAEF by replacing the histogram match component with other statistical indices, i.e. kurtosis and earth mover's distance, or by adding a fourth or fifth component such as kurtosis and skewness. The existing spatial metrics i.e. SPAtial Efficiency (SPAEF), Structural Similarity (SSIM) and Spatial Pattern Efficiency Metric (SPEM) were compared with newly proposed metrics to assess their converging performance. The mesoscale Hydrologic Model (mHM) of the Moselle River is used to simulate streamflow (Q) and actual evapotranspiration (AET). The two-source energy balance (TSEB) AET during the growing season is used as monthly reference maps to calculate the spatial performance of the model. The Moderate Resolution Imaging Spectroradiometer (MODIS) based Leaf area index (LAI) is utilized by the mHM via pedo-transfer functions and multi-scale parameter regionalization approach to scale the potential ET. In addition to the real monthly AET maps, we also tested these metrics using a synthetic true AET map simulated with a known parameter set for a randomly selected day. The results demonstrate that the newly developed four-component metric i.e. SPAtial Hybrid 4 (SPAH4) slightly outperform conventional three-component metric i.e. SPAEF (3% better). However, SPAH4 significantly outperforms the other existing metrics i.e. 40% better than SSIM and 50% better than SPEM. We believe that other fields such as remote sensing, change detection, function space optimization and image processing can also benefit from SPAH4.

**Keywords:** mHM, model calibration, spatial pattern, SPAEF, MODIS, TSEB

## 1 Introduction

Distributed hydrologic models have a crucial role in creating digital twin of the water cycle in nature by revealing physical mechanisms and process interactions. After identifying the best parameter set through calibration, these models are used to conduct robust numerical experiments assessing climate change impacts (Beven, 2023) or land use land cover change impacts on model output fluxes such as runoff (Busari et al., 2021), groundwater recharge, soil moisture and actual evapotranspiration (AET). A skillful model enables decision-makers to plan for and respond to water-related extremes such as hydrological droughts and floods. Accuracy of the model results depends on the success of identifying best combination of the parameters since calibration process helps us reduce discrepancies in model physics. Demirel et al. (Demirel et al., 2018) showed that using only streamflow hydrograph performance as objective function diminishes the AET patterns simulated by the model. However, incorporating satellite based remotely sensed AET into the multi-objective calibration framework that has already streamflow, surprisingly improves both water balance and AET performance of the model. Other studies benefitted from land surface temperature (Zink et al., 2018), soil moisture (López et al., 2017; Wakigari & Leconte, 2023), AET (Avcuoğlu & Demirel, 2022; Gaur et al., 2022; Odusanya et al., 2022; Sirisena et al., 2020) and groundwater (Danapour et al., 2021; Stisen et al., 2018) in hydrologic model calibration.

In other words, hydrologic model calibration is essential for ensuring the validity and reliability of model predictions i.e. of most important for water management and decision-making processes. However, the robustness of hydrologic model calibration heavily relies on how the model is guided in the solution space via the performance metrics (de Boer-Euser et al., 2017; Knoben et al., 2019; Martinez-Villalobos et al., 2022; Onyutha, 2022; Schneider et al., 2022). If the metric is too loose (tolerant) or prone to the sampling uncertainty (Clark et al., 2021), the calibration process can stop quickly in the local minima while the modeler searches for the best global solution. The key point of the modelling chain is the selection of appropriate metric. Our study focuses on development of a novel metric with least tolerance (highest discrimination skill) based on benchmarking existing metrics in evaluating the similarity of two raster maps. We are particularly interested in multi-component bias-insensitive spatial metrics for pattern comparison. Thus, bias sensitive temporal metrics used for water balance are not within the scope of this study.

The use of multi-component spatial metrics in hydrologic model calibration is an important advancement in the field of water resource management and resource allocation. The multi-component metrics provides a more nuanced evaluation of model performance compared to traditional single-component metrics e.g. mean absolute error and coefficient of determination. The adoption of these metrics allows for a more comprehensive understanding of the hydrologic system and its spatial variability, which is critical for informed decision-making. These metrics differ from single-component metrics in that they consider multiple components of the hydrological system, rather than just one component. By providing a more comprehensive evaluation of the hydrologic system, multi-component metrics help to identify areas where models can be improved. For spatial metrics, the added level of complexity provided by multi-component metrics offers a more robust evaluation of model performance, providing a better understanding of the spatial variability of the hydrologic system.

In recent years, remote sensing data from satellites, such as Moderate Resolution Imaging Spectroradiometer (MODIS) products, have become commonly used in hydrologic model calibration since this product provides estimates of AET from vegetation, which is a key component and major water loss in the hydrologic cycle (Becker et al., 2019; Rientjes et al., 2013). On one hand, it serves to better represent the cell-to-cell hydrological dynamics and

diversity in the basin also allows for a more detailed understanding of the water budget at the land surface and helps to better quantify the water requirements of vegetation. On the other hand, the MODIS Leaf Area Index (LAI), product provides information about the leaf area index, which is a measure of the amount of vegetation cover in an area. This information is essential for understanding how vegetation influences the water cycle by affecting factors such as precipitation, evapotranspiration, and runoff. In this study, we use LAI to dynamically scale the PET input to the model to improve AET performance and present a comprehensive benchmarking of multi-component spatial metrics using MODIS-LAI and TSEB AET products, to assess their potential for calibration (Immerzeel & Droogers, 2008).

There are various performance metrics in hydrology. The Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) are the most widely recognized performance metrics used in evaluating and calibrating rainfall-runoff models. These two metrics have been instrumental in advancing our understanding of hydrological processes and improving the performance of hydrologic models (Gupta et al., 2009; Nash & Sutcliffe, 1970). They have paved the way for the development of more advanced and sophisticated performance evaluation techniques. Despite the sampling uncertainty inherited in these metrics (Clark et al., 2021), NSE and KGE continue to be widely accepted in the hydrology community due to their simplicity and effectiveness in evaluating model performance. Many of the newer metrics that have been introduced in recent years have been inspired by and built upon the foundation established by NSE and KGE. The conventional model calibration relies on using flow-oriented temporal metrics, such as the NSE and KGE. However, these metrics have a limitation as they lack spatial considerations and are prone to the sampling uncertainty. This has driven the need for development of intolerant spatial performance metrics which can better evaluate and improve the spatial accuracy of a hydrologic model. Spatial-pattern-oriented SPAtial Efficiency (SPAEF) metric developed by Demirel et al. (Demirel et al., 2018) builds upon the strength of KGE and incorporates new idea of distribution comparison via histogram overlap index. It is designed as a multi-component metric specifically suited for comparing spatial patterns of two raster maps, with its three main data properties being co-location, variation, and distribution. Although SPAEF was primarily developed for hydrologic community, it has been used in many different disciplines such as atmospheric circulation modeling (Ahmed et al., 2019), flood risk analysis (Hossain & Meng, 2020), function space optimization, fisheries (Thoya et al., 2021) and neuroscience (Yoo et al., 2020). In these studies, SPAEF has been tested and proven to be robust and easy to interpret due to its three distinct and complementary components of correlation, variance and histogram matching. Following the multi-component structure idea, we present new metrics in this study to improve SPAEF by adding fourth of fifth new components or replacing histogram match with other components. Using this approach, we aimed for reducing uncertainty in the new metric and make it sharp (discriminant) when evaluating patterns on two raster maps whether they are similar or not.

In recent literature, there has been attempts to revise SPAEF component i.e. Spatial Pattern Efficiency Metric (SPEM) (Dembélé et al., 2020). Similar to SPAEF, it has been proposed as a bias-insensitive and multi-component spatial pattern-oriented metric using satellite remote sensing data. Structural Similarity index (SSIM) is another pattern-oriented metric, it stands out with its spatial structure (Nilsson & Akenine-Möller, 2020; Wang et al., 2004). It was proposed by Wang et al. (Wang et al., 2004) for image quality assessment and has been used in different studies such as medical imaging, ecological restoration, and change detection in the hydrological cycles and remote sensing images (Arun et al., 2021; Dougherty et al., 2020; Wiederholt et al., 2019). Knoben et al. (Knoben et al., 2019) compared NSE and KGE metrics and argued that instead of relying directly on the KGE value, the components should be analyzed in depth, even the weighting of the components. A study analyzing sampling

uncertainty in popular performance metrics in hydrologic modeling highlighted that the KGE can be heavily influenced by just a few data points (Clark et al., 2021). A study on the hydrological model skill score compared metrics with different forms of correlation and measures of variability, claiming the term covariance is more appropriate for evaluation (Onyutha, 2022). Another recent study, based on the largest residuals, focused on reducing the largest errors, and argued that metrics should be less sensitive to errors and more sensitive to bias (Schneider et al., 2022). The publication (Martinez-Villalobos et al., 2022) compared metrics for evaluating precipitation probability distributions by comparing climate model simulation data with real platform satellite data, therefore they showed the importance of probability distribution functions. A study from the Netherlands (de Boer-Euser et al., 2017) stated that strong components can be included in different metrics rather than considering a single general metric for model comparison.

The existing spatial metrics aimed for the best convergence using terms such as correlation, variation, histogram intersection, and root mean square error. However, kurtosis has hitherto been an underrated term for spatial performance, and a four-component spatial-pattern-oriented metric also does not exist for the hydrologic model calibration. We used the kurtosis ratio by including it as a new component for the first time in this study in order to achieve the best spatial convergence and fit. With the addition of a new component, the weighting by which the components affect the value has also changed. By revealing the effect of kurtosis on spatial performance, we developed a new four-component metric that does not require user input.

We aim to investigate the best potential to use multi-component spatial metrics in hydrological model calibration, by proposing a new multi-component spatial metric that especially includes the kurtosis component and benchmarking it to existing multi-component spatial metrics. The primary purpose of this study is to evaluate the performance of the hydrological model using multicomponent spatial metrics and to determine the potential impact on model accuracy and precision. In addition, this study aims to identify the most effective combination of spatial metrics for hydrological model calibration and to develop a framework for future work in this area. A large number of metrics in the literature creates confusion and difficulty for users to choose from, so we compared metrics in this study to look for the most successful one to put a stop to metric redundancy. Addressing these goals, this study aims to contribute to ongoing research efforts to improve the accuracy and reliability of hydrological models.

The accuracy of the analysis has been increased by comparing model predictions with real platforms. It is aimed to improve the convergence between observed and simulated maps by using two-source energy balance (TSEB) model's AET data. The MODIS-LAI data were used both to correct the PET and to represent the vegetation dynamics of the Moselle basin. We utilize a spatially distributed mesoscale Hydrologic Model (mHM) with it features pedo-transfer functions for LAI data and a Multiscale Parameter Regionalization (MPR) approach to scale the potential ET (Kumar et al., 2013; Samaniego et al., 2010). We tested our framework in three different cases to provide comprehensive outlook to the calibrations i.e. 100 iterations were applied in the first case and 1000 iterations in the second case, so the effect of the number of iterations was also assessed. In the third case, reproducibility was achieved by analyzing the randomly selected synthetic map. OSTRICH software (L. Shawn Matott, 2004; L.S. Matott, 2017) was used as the calibration tool and Parallel Dynamically Dimensioned Search Algorithm (PDDS) was used as the calibration algorithm (Asadzadeh & Tolson, 2013). The combined SPAEF value of the growing season was used as the main objective function for ET, and the KGE was presented for discharge (Q) in addition. We developed multiple metrics with different components and different component numbers, trying to increase the effectiveness (sharpness) of each component on convergence performance. We made an elaborated

comparison between the existing performance metrics in the literature and the newly developed metrics based on SPAEF. As a result of the rigorous assessment of metrics, we identified not only the superior but also new metric. The strongest aspect of this new metric is the added kurtosis component.

## 2 Study area and data

### 2.1 Study area

The study area is the Moselle River basin, the largest part of the Rhine River basin, of which it is one of the main tributaries, characterized by diverse landforms (**Figure *1***). The origin of the river from the Vosges Mountains before the interterritorial transfer from France to enter Germany and Luxembourg. Furthermore, at the triangle where Germany, France and Luxembourg meet, the Moselle River becomes the borderline between Germany and Luxembourg for 36 km. Also, it has a surface area of approximately 27262 km$^2$ and a length of 545 km. Whereas, land use in the basin includes forestry, agriculture and cattle breeding in the mountains and hillslopes, winegrowing on vineyards of sunny valley slopes. Moreover, the altitude varies from 59 to 1326 m, with an average altitude of around 340 m (Demirel et al., 2013). In addition to having 26 sub-basins with surface areas varying from 102 to 3353 km$^2$, the river flow is organized by different dams, dikes, powerplants and locks such as the Trier Dam, Koblenz Dam and Detzem Lock. The outlet discharge at Cochem station, located between Trier and Koblenz, varies from 14 m$^3$/s in dry summers to a maximum of 4000 m$^3$/s during winter floods, with a mean discharge of around 315 m$^3$/s (Demirel et al., 2015).

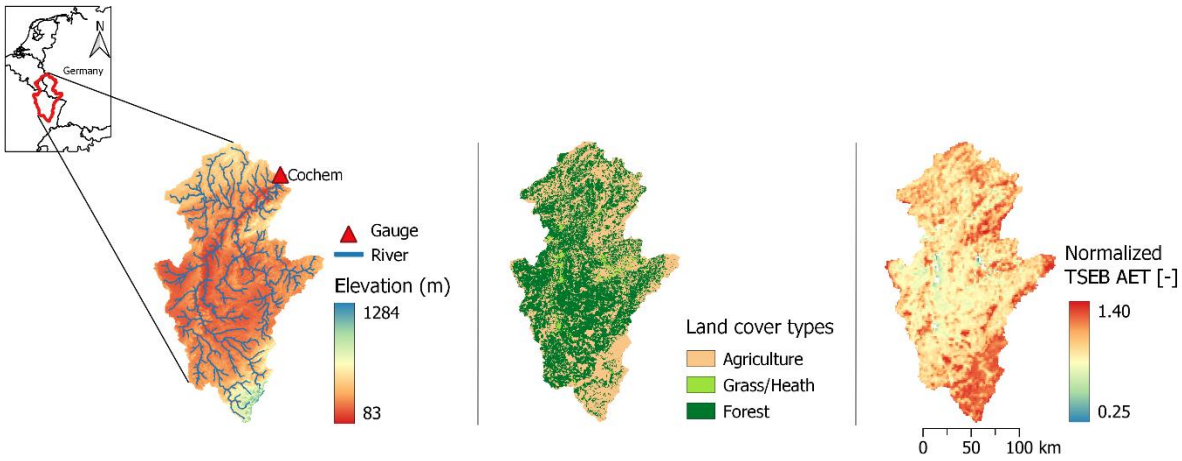



**Figure 1.** DEM, land cover and AET characteristics of Mosel River basin.

An average pattern of satellite-based actual evapotranspiration for July (average of all years from 2002 to 2014) is presented to illustrate the interaction between DEM and land cover characteristics that generate the land surface flux patterns.

### 2.2 Satellite data

MODIS has a vital role in obtaining the satellite-based data used in this study, is an essential sensor aboard the Terra (EOS AM) and Aqua (EOS PM) satellites for the earth and climate measurements at a spatial resolution of approximately 1 km × 1 km. It provides terrestrial, atmospheric and thalassic data and a view of the entire Earth's surface for large and diverse user communities around the world. In this study, TSEB based AET is used as reference spatial

patterns (Allen et al., 1998; Norman et al., 1995). TSEB is an energy balance model using the energy flux principle by separating into two-layer, vegetation and soil.

The water limited growing season was chosen as the analysis period because it avoids climate gradient on the AET patterns emphasizing vegetation dynamics instead of wet soil conditions i.e. AET that is equal to the PET. All remote-sensing-based AET data were converted to long term monthly mean data during the growing season across all years for the model calibration period (2002–2014). In what follows, three-monthly mean periods were obtained with a total of three-term between March and November, i.e. March-April-May (MAM), June-July-August (JJA), and September-October-November (SON), representing AET under cloud-free conditions. We will attribute these AET maps as reference observations, although they are estimates from an energy balance model based on satellite observations and not pure observations.

**Table 1.** Overview of morphological and meteorological data used as input for mHM (Rakovec et al., 2016).

| Variable | Description | Spatial resolution (degrees) | Source |
|---|---|---|---|
| Q (daily) | Streamflow | Point | GRDC |
| P (daily) | Precipitation | 0.0625 | E-OBS |
| PET (daily) | Potential evapotranspiration based on Hargreaves and Samani (Hargreaves & Samani, 1985) | 0.0625 | E-OBS |
| $T_{avg}$ | Average air temperature | 0.0625 | E-OBS |
| LAI | Fully distributed 12-monthly values based on 8-day time-varying leaf area index (LAI) dataset | 0.001953125 | MODIS |
| Land cover | Forest, agriculture and urban | 0.001953125 | MODIS |
| DEM-related data | Slope, aspect, flow accumulation and direction | 0.001953125 | SRTM |
| Geology class | Two main geological formations | 0.001953125 | ESD – UFZ Leipzig (Rakovec et al., 2016) |
| Soil class | Fully distributed soil texture data | 0.001953125 | HWSD |

GRDC – Global Runoff Data Centre, E-OBS – The gridded observational dataset from Copernicus, MODIS – Moderate Resolution Imaging Spectroradiometer, SRTM – Shuttle Radar Topography Mission, ESD – European Soil Database, HWSD – Harmonized World Soil Database

## 3 Hydrological model

This research utilizes the mesoscale Hydrologic Model (mHM) v.5.11.2 (Samaniego et al., 2021) which is a grid-based spatially distributed model it features pedo-transfer functions and MPR (Kumar et al., 2013; Samaniego et al., 2010; Thober et al., 2019). Another feature of mHM is the use of leaf area index (LAI) data not only for calculating interception loss but also for dynamically scaling PET (Demirel et al., 2018). With these unique features, it is more flexible than other existing hydrologic models in line with the purpose of this study. The model features 69 adjustable global parameters that can be optimized during the calibration process (Demirel et al., 2018). The model works on the basis of water balance rather than energy balance and provides various physically meaningful spatial outputs, fluxes and states as simulating major elements of the hydrologic processes, i.e. soil moisture dynamics, interception, infiltration, evapotranspiration, snow accumulation and melting, groundwater storage, seepage, surface runoff and others.

The basic data for the running mHM can be classified into meteorological data, morphological data, land cover data and gauge streamflow data. Table 1 shows a summary of the data used in mHM setup provided by Rakovec et al. (Rakovec et al., 2016). As seen in the table, mHM can handle different spatial resolutions of meteorological data and morphological data since it has internal upscaling and downscaling subroutines. At this point, the Multi-Scale Parameter Regionalization technique comes into play and enables user to map calibrated parameters to the simulated grids with pedo-transfer functions. This approach prevents uniform parameter fields and protects sub-grid heterogeneity of the fluxes. In other models, every parameter gets the same value in the entire sub-basin or in each hydrologic response units resulting in uniform flux results for the same domain.

The meteorological model inputs are precipitation, average air temperature and potential evapotranspiration (PET). In our study, PET was direct input to the mHM and estimated outside with Hargreaves-Samani (Hargreaves & Samani, 1985) method using additional temperature data. All meteorological data are obtained from E-OBS at daily resolution, originally at 10-20 km. The morphological variables are digital elevation model (DEM), soil maps with textural features, geological maps including specific yield, permeability and aquifer thickness. In addition to characterizing the morphology of the basin, DEM masks the grid cells with the basin boundaries to eliminate no-data parts. All morphological data are prepared at 0.001953125 degrees (~200 m × 200 m) scale. The model hydrology is evaluated at 0.015625 degrees (~2x2 km) spatial resolution and daily time step. Lastly, monthly leaf area index (LAI) maps are used to represent the vegetation dynamics for both interception calculation and PET correction for the entire period (2002–2014). Four years of model warm-up period (1998–2001) is used. Observed daily streamflow (Q) data at Cochem (station #6336050), provided by the Global Runoff Data Centre (GRDC), Koblenz (Germany), is used to calibrate water balance in the basin.

## 4 Methods

In this study, we tested nine different spatial metrics i.e. two of them are existing metrics, and seven of them are newly developed based on SPAEF (Table 2). To evaluate the effect of number of iterations, calibrations were pursued with either 100 or 1000 maximum iterations. Besides, synthetically created AET maps using mHM and a pre-defined parameter set are utilized to mimic a “hide and seek” case. This is crucial to test the guidance performance of the metrics in the multi-dimensional solution space to find the hided (perfect) solution within 1000 iterations since search algorithms, i.e. ParaPADDS algorithm herein, require a metric to evaluate model results at every iteration.

### 4.1 Objective Functions

Multi-component structure of our metrics was inspired by the Kling–Gupta efficiency (Gupta et al., 2009). KGE is one of the most used metrics in the hydrologic modelling to evaluate streamflow performance. As shown in Eq.(1), it has three components, i.e., correlation, variability and bias.

$$\mathrm{KGE} = 1 - \sqrt{\left(\alpha_Q - 1\right)^2 + \left(\beta_Q - 1\right)^2 + \left(\gamma_Q - 1\right)^2}$$
$$\alpha_Q = \rho(o, s), \beta_Q = \frac{\sigma_S}{\sigma_O} \; and \; \gamma_Q = \frac{\mu_s}{\mu_O} \tag{1}$$

where $\alpha_Q$ is the Pearson correlation coefficient between the observed (o) and the simulated (s) discharge time series, $\beta_Q$ is the relative variability based on the ratio of standard deviation in simulated and observed values and $\gamma_Q$ is the bias fraction which is normalized by the standard deviation of the observed data.

Table 2 shows the summary of SPAEF based metrics. For brevity, we used Eq. (2) as formula template i.e. a generic formulation type that encompasses in the number and content of components. The excessed style in Eq (2) includes all metrics form with various components.

$$\mathrm{METRIC} = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2 + (\kappa - 1)^2 + (\delta - 1)^2} \tag{2}$$

**Table 2.** SPAEF based metrics used as objective functions.

| Metric | Components | | | | | Index |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\gamma$ | $\kappa$ | $\delta$ | |
| SPAtial Efficiency (**SPAEF**) | $\rho(o,s)$ | $\frac{\sigma_o}{\mu_o} / \frac{\sigma_s}{\mu_s}$ | $\frac{\sum_{j=1}^{n} min(K_j,L_j)}{\sum_{j=1}^{n} K_j}$, n=100 fixed | none | none | Eq. (3) |
| SPAtial EFficiency Prime (**SPAEF'**) | same as SPAEF | same as SPAEF | same as SPAEF except for dynamic $n$ i.e. number of bins $n = floor\{\sqrt{length(o)}\}$ | none | none | Eq. (4) |
| SPAtial Count Density Efficiency (**SPACD**) | same as SPAEF | same as SPAEF | $\frac{\sum_{j=1}^{n} min(K_j,L_j)}{\sum_{j=1}^{n} K_j}$ $(v_n = c_n/w_n)$ | none | none | Eq. (5) |
| SPAtial Hybrid 4 Efficiency (**SPAH4**) | same as SPAEF | same as SPAEF | same as SPAEF' | $\frac{Kurt(s)}{Kurt(o)}$ | none | Eq. (6) |
| SPAtial Kurtosis Efficiency (**SPAK**) | same as SPAEF | same as SPAEF | none | same as SPAH4 | none | Eq. (7) |
| SPAtial Hybrid 5 Efficiency (**SPAH5**) | same as SPAEF | same as SPAEF | same as SPAEF' | same as SPAH4 | $\frac{Skew(s)}{Skew(o)}$ | Eq. (8) |
| SPAtial Histogram Equalization Efficiency (**SPAHE**) | same as SPAEF | same as SPAEF | $\frac{\sum_{j=1}^{n} min(K_j,L_j)}{\sum_{j=1}^{n} K_j}$ | none | none | Eq. (9) |
| SPAtial Movers' Distance Efficiency (**SPAMD**) | same as SPAEF | same as SPAEF | $\frac{\sum_{i=1}^{K}\sum_{i=1}^{L} f_{i,j} d_{i,j}}{\sum_{i=1}^{K}\sum_{i=1}^{L} f_{i,j}}$ | none | none | Eq. (10) |
| Spatial Pattern Efficiency Metric (**SPEM**) | $1 - \frac{6\sum_{1}^{n} d^2}{n(n^2-1)}$ | same as SPAEF | $1 - E_{RMS}(Z_{X_s}, Z_{X_o})$ | none | none | Eq. (11) |

SPAEF is the seed of our newly proposed metrics as our aim is to sharpen SPAEF. In other words, we intend to improve its discriminating power while judging whether two maps are similar or not. SPAEF uses a multi-component structure of the KGE metric. In Eq. (3), $\alpha$ is the Pearson correlation coefficient between the observed (o) and simulated (s) pattern, $\beta$ is the fraction of the coefficient of variation representing spatial variability and $\gamma$ is the histogram intersection, which based on z-scores, for the given histogram K of the observed pattern and the histogram L of the simulated pattern, each containing n bins (Swain & Ballard, 1991). The SPAEF can have a value between $-\infty$ and 1, where a value closer to 1 indicates highest spatial similarity between the observations and model simulations (Koch et al., 2018).

As a result of various adjustments and improvements made in the SPAEF components, new metrics were proposed and tested i.e. SPAEF', SPACD, SPAH4, SPAK, SPAH5, SPAMD, and SPAHE. We included two popular metrics, SPEM and SSIM into benchmark.

First improvement in SPAEF is changing user defined the number of bins to an automated n based on the number of elements (grids) in the raster map (see Eq (3)). We introduced a simple approach i.e. the square root of the length of the observed data as $n = floor\{\sqrt{length(o)}\}$ although there are different methods for the same purpose (Freedman & Diaconis, 1981; Scott, 1979; Sturges, 1926). This slightly new version of the SPAEF is presented as SPAEF-Prime (SPAEF') as shown in Eq (4). Unlike the standard version, the SPAEF' does not require any user-defined inputs now.

Eq (5) shows Spatial Count Density Efficiency (SPACD) which has a different type of normalization based on count density approach in the calculation of the histogram intersection component. While the first two components remain constant as in SPAEF' the calculation of n in the gamma component has changed. This approach uses count or frequency scaled by the width of the bin $v_n = c_n \,/\, w_n$, $v_n$ is the bin value, $c_n$ is the number of elements in the bin and $w_n$ is the width of the bin, respectively.

Eq (6) shows SPAtial Hybrid 4 Efficiency (SPAH4) which is a four-component metric obtained by adding kurtosis i.e. a fundamental statistical property of distributions to the SPAEF' metric. Kurtosis can be defined as a measure of how prone a distribution is to outliers (Pearson, 1905). SPAH4 offers a more accurate perspective by questioning not only the match of the histograms but also the extreme values and spread in the data. The 4$^{th}$ component is symbolized by the expression $Kurt$ and $\kappa$ is the ratio of the kurtosis coefficients of the simulated (s) and observed (o) data. Eq. (7) shows SPAtial Kurtosis Efficiency (SPAK) which is a three-component metric replacing the histogram intersection component in the SPAEF metric with the kurtosis coefficient component. Thus, it dominates the metric on its affinity for discrete values without questioning histogram intersection. $\alpha$ and $\beta$ were introduced and explained in previous metrics, also $\kappa$ is declared in Eq. (7) as ratio of kurtosis coefficient. This metric can be characterized as a mixture of SPAH4 and SPAEF metrics. Eq. (8) shows SPAtial Hybrid 5 Efficiency (SPAH5) which is a five-component metric adding skewness to the SPAH4 metric. Skewness can be defined as a measure of the asymmetry of the data around the sample mean.

Eq. (9) shows SPAtial Histogram Equalization Efficiency (SPAHE) that is very similar to SPAEF with additional step before histogram match calculation "histogram equalization" approach. This approach is a computer image processing technique used to improve contrast in raster data. Its quantitative logic is based on the grayscale transformation ($T$) to minimize $\left|c_1\big(T(k)\big) - c_0(k)\right|$, $c_0$ is the cumulative histogram of the input data, and $c_1$ is the cumulative sum of target histogram for all intensities $k$. Histogram equalization is a specific case of the histogram remapping methods. It is an image processing technique used to advance contrast in

images which spatial patterns for this study. It achieves this by efficaciously sprawling out the most frequent intensity values, i.e. expanding the intensity range of the image (Efford, 2000).

Eq. (10) shows SPAtial Efficiency Movers' Distance (SPAMD) is another SPAEF-oriented multi-variate metric which measures the quantitative closeness of two pattern set by considering the Earth Movers' Distance of their histograms (Rubner et al., 1998). The aim of EMD approach is minimization of overall transfer cost in the conversion one histograms to another. In Eq (10), $f_{i,j}$ is flow cost of transfer ith term of histogram K of observed map to jth histogram L simulated map at distance $d_{i,j}$. EMD is the ratio of work done through the total optimal flow and the total flow. The value of EMD is zero indicates the perfect consistency between two histograms.

Eq. (11) shows Spatial Pattern Efficiency Metric (SPEM), a metric inspired by KGE and SPAEF, is one of the existing metrics included in our analysis (Dembélé et al., 2020). It forces the z-scores of simulated variables and observed variables to be equal (i.e., minimizing their ERMS) corresponds to matching their grid cell locations (i.e., spatial patterns). SPEM considers a modeled variable (Xmod) and an observed variable (Xobs) of n elements, it is defined as Eq. (11); where *rs* is the Spearman rank-order correlation coefficient with $d$ the difference between the ranks of Xmod and Xobs. γ is the variability ratio that assesses the similarity in the dispersion of the probability distributions of Xmod and Xobs, with μ and σ representing the mean and the standard deviation, respectively, and α the spatial location matching term calculated as the root-mean-square error (ERMS) of the standardized values (z-scores, ZX) of Xmod and Xobs (Dembélé et al., 2020). The formula for $d$ can be written as $d = diff\big(rank(X_s), rank(X_o)\big)$. SPEM ranges from $-\infty$ to 1, which is its optimal value.

Lastly, Eq. (12) shows Structural Similarity index (SSIM) (Wang et al., 2004). An image quality metric SSIM to evaluate degradation grade caused by visual data processing. This method considers pattern similarity as it detects changes in the variation of structural information between the two images. The algorithm formulates perception sensibility to visual changes based on the distortion luminance, contrast and structure information. By combining three components, similarity can be characterized with overall unit metric in terms of statistical properties of simulated and observed data such as mean $\mu$, standard deviation $\sigma$ and covariance $cov_{o,s}$, as shown in Eq. (12). $c_1$, $c_2$ are constants that stabiles functions when the dominator terms are close to zero. The SSIM is a fully referenced objective quality metric that gives values in the range [0,1] relative to the structural relationship between the two images.

$$\boldsymbol{SSIM} = \frac{(2\mu_o\mu_s + c_1)\left(2cov_{o,s} + c_2\right)}{({\mu_o}^2 + {\mu_s}^2 + c_1)\,({\sigma_o}^2 + {\sigma_s}^2 + c_2)} \tag{12}$$

All nine spatial metrics were calculated separately as long term (2002-2014) monthly average of AET data for three periods covering the growing season and combined as in Eq (13) to minimize the total error, representing objective function (OF). These periods are symbolized as March-April-May (MAM), June-July-August (JJA), and September-October-November (SON).

$$\boldsymbol{Minimize}\ [(1 - METRIC_{MAM})^2 + \left(1 - METRIC_{JJA}\right)^2 + (1 - METRIC_{SON})^2] \tag{13}$$

It should be noted that although we tested other metrics and approaches, we only reported nine selected metrics in this study. For instance, we used harmonic mean or geometric mean instead of the arithmetic mean in the second component of SPAEF. In another attempt, we replaced

the skewness coefficient ratio with different L-moments. We also used Hausdorff distance (Hausdorff, 1914) and Fréchet distance (Fréchet, 1906) as third component in SPAEF. Even we used the product of components i.e. multiplied them instead of adding them. However, all these attempts did not reveal better results than those reported in this study. Therefore, for brevity we reported the ranking of only these nine metrics above. In this calibration study, we fine-tuned only 20 parameters of daily mHM for the Mosel Basin using the popular global search algorithm Pareto-Archived Dynamically Dimensioned Search (ParaPADDS) algorithm (Asadzadeh & Tolson, 2013) using 750 maximum iteration and 3 parallel cores. The 20 parameters out of 69 mHM parameters are selected based on a sensitivity analysis done in our previous study. Note that ParaPADDS is the multi-objective version of the Dynamically Dimension Search algorithm (Tolson & Shoemaker, 2007) available in OSTRICH Optimization Software Toolkit (L.S. Matott, 2017).

## 5 Results

In this study, six novel metrics are proposed and compared with existing SPAEF, SPEM and SSIM metrics in pattern analysis of distributed hydrologic model simulations. The new metrics can be called as "the sisters of SPAEF" as they have emerged from the well-established SPAEF with additional unique statistical features such as automated number of bins, kurtosis and skewness included in their structure. We ranked the nine metrics based on their effectiveness in distinguishing between two raster maps during distributed model calibration with MODIS-LAI and TSEB AET for a period of 13 years from 2002 to 2014. Pre-selected 20 mHM parameters are included in the following three different pattern-only calibration cases: (1) 100 iterations with satellite data, (2) 1000 iterations with satellite data, and (3) 1000 iterations with synthetic maps. Synthetic map represents a map simulated with a known mHM parameter set for a randomly selected day that is used as the target in parameter optimization (calibration) process. The use of this synthetic scenario is planned to ensure the reproducibility of the analysis and to have a fully controlled numerical experiment. Obviously, long term monthly averaging was done only with real satellite data to form robust seasonal pattern maps i.e. target in the calibration.

Although water balance metrics, i.e. temporal metrics, are not included in the calibration, KGE values are calculated to evaluate the model simulations together with standard SPAEF in Table 3. Streamflow simulation performance was calculated for the calibration period (2002-2014), using the KGE metric between the observed gauge streamflow and simulated streamflow from the model. This is done only for case 1 (TSEB 100 runs) and 2 (TSEB 1000runs) i.e. real satellite data are used in the pattern-based optimization. It is interesting to note that some of the pattern metrics help to improve the bias in water balance as well. The three OF columns in this table show lowest (best) values of each metrics reached using Eq. (13). This is particularly important to show the skill of the nine metrics in converging to zero i.e. certainly exists in the synthetic case (3). It should be noted that the metrics are ranked based on the standard SPAEF values. Closer inspection of the Table 3 shows that TSEB 1000 iterations significantly improves the SPAH4 performance from 0.608 to 0.688 (SPAEF value) as compared to the TSEB 100 iterations. The reduction in OF is even more remarkable since the error in SPAH4 was halved from 0.70 to 0.35 when iterations are increased to 1000. It is clear from this table that SPAHE and SPAH5 are the worst performing two metrics among all three cases.

Comparing the two results (100 runs vs 1000 runs) it can be seen that all metrics are improved with the increased number of iterations showing the importance of the selecting appropriate number of the iterations for the search algorithm. However, if enough freedom is not given to the optimizer, it may fail to find the global optimum point in the solution space. Combining kurtosis with skewness in the same metric (SPAH5) did not produce a discriminative metric.

This result is somewhat counterintuitive as we expect more constrain would yield improved performance. What is striking about the values in this table is histogram equalization step did not help to improve the pattern results and discriminative power of the metric.

**Table 3.** Calibration results of the three cases. Note that metrics are ranked based on 1000 run - SPAEF values (4th numeric column).

| Metrics | TSEB 100 runs | | | TSEB 1000 runs | | | SYNTHETIC MAP 1000 runs | |
|---|---|---|---|---|---|---|---|---|
| | SPAEF | KGE | OF | SPAEF | KGE | OF | SPAEF | OF |
| **SPAH4** | 0.608 | 0.78 | 0.70 | **0.688** | 0.77 | 0.35 | 0.948 | 0.05 |
| **SPACD** | 0.619 | 0.26 | 0.40 | 0.673 | 0.74 | 0.27 | 0.939 | 0.04 |
| **SPAEF'** | 0.585 | 0.36 | 0.52 | 0.671 | 0.52 | 0.33 | 0.949 | 0.05 |
| **SPAK** | 0.558 | 0.89 | 0.39 | 0.638 | 0.87 | 0.25 | 0.906 | 0.01 |
| **SPAMD** | 0.614 | 0.07 | 0.29 | 0.625 | 0.66 | 0.21 | 0.859 | 0.02 |
| **SSIM** | 0.557 | 0.21 | 0.19 | 0.491 | 0.41 | 0.15 | 0.948 | 0.00 |
| **SPEM** | 0.609 | 0.33 | 1,71 | 0.460 | 0.61 | 1,46 | 0.941 | 0.05 |
| **SPAHE** | 0.492 | 0.70 | 0.25 | 0.376 | 0.65 | 0.21 | 0.758 | 0.04 |
| **SPAH5** | -0.519 | 0.61 | 8,15 | 0.211 | 0.53 | 2,07 | 0.953 | 0.05 |

What stands out in the table is that SSIM seems to be the most tolerant metric reaching lowest OF values which corresponds to the poor SPAEF performance in all three cases. In case 3, in particular, the search algorithm could converge nearly to zero SSIM but the evaluation of the maps with SPAEF revealed that it is only a match around 0.95 SPAEF and not very close to 1 SPAEF i.e. perfect pattern match. In other words, minimizing SSIM in Eq (13) nearly to zero after calibration doesn't guarantee a perfect pattern match in terms of SPAEF currency (metric). Based on the results of case 1 and 2, SPAH4 and SPAK are the most successful spatial metrics for water balance. Obviously, SSIM and SPAMD have the worst KGE performance in case 1 and 2. Note that KGE is not calculated for the synthetic case 3. Interestingly, the minimization of SPEM and SPAH5 metrics via Eq (13) after optimization resulted in poor values above 1 both in case 1 and 2.

Figure 2 shows the reference AET maps and simulated AET maps from the mHM with calibrated parameters after 100 iterations (case 1). The reference three maps are given in both columns for ease of comparison. The order of the metrics is in accordance with the performance ranking in Table 3 and also, the ranking is provided (e.g. #1, #2 etc.) to help to the reader. The combined SPAEF values of three periods (MAM, JJA and SON) are presented in brackets underneath the metric name. To use a single legend, the maps are normalized with their mean. The resultant maps from SPACD and SPAMD (second row in Figure 2) are slightly better than other rows as visually more similar to the reference maps (first row in Figure 2). Closer inspection of the maps shows that the high contrast between west and south of the basin in SON period is well-captured by most of the metrics except for the SPAH5 (row 6, rank #9).

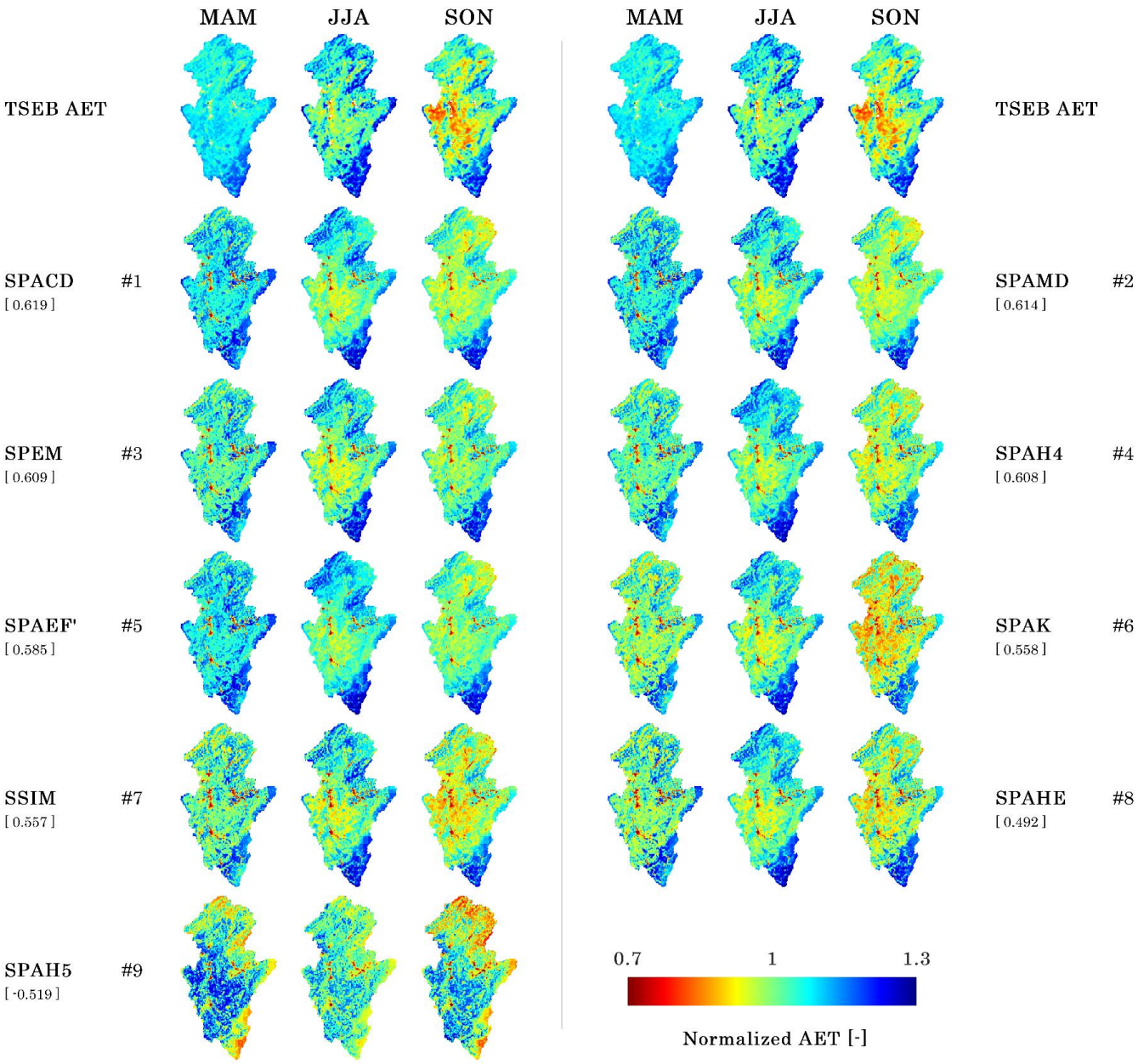



**Figure 2.** Long term average three-monthly TSEB reference maps versus mHM simulated maps using MODIS-LAI and best-balanced Pareto solution parameter set from 100 run case.

Figure 3 shows the reference AET maps and simulated AET maps from the mHM with calibrated parameters after 1000 iterations (case 2). It is consistent with Figure 2 that the simulated AET maps by the model parameter sets optimized with SPAH4 and SPACD metrics are most close to the reference maps. Similarly, the poor AET performance of SPAH5 maps is apparent from the maps in the last row of the figure. Map illustration of each period reveals that the combined metric value (OF) can hinder individual map performance. For instance, the SON map of the SPAHE metric in Figure 3 shows that the model better converges to the remotely sensed reference map when optimized with SPAHE whereas the MAM and JJA maps show that the model could not reproduce the AET maps of these periods as successful as with the other metrics.

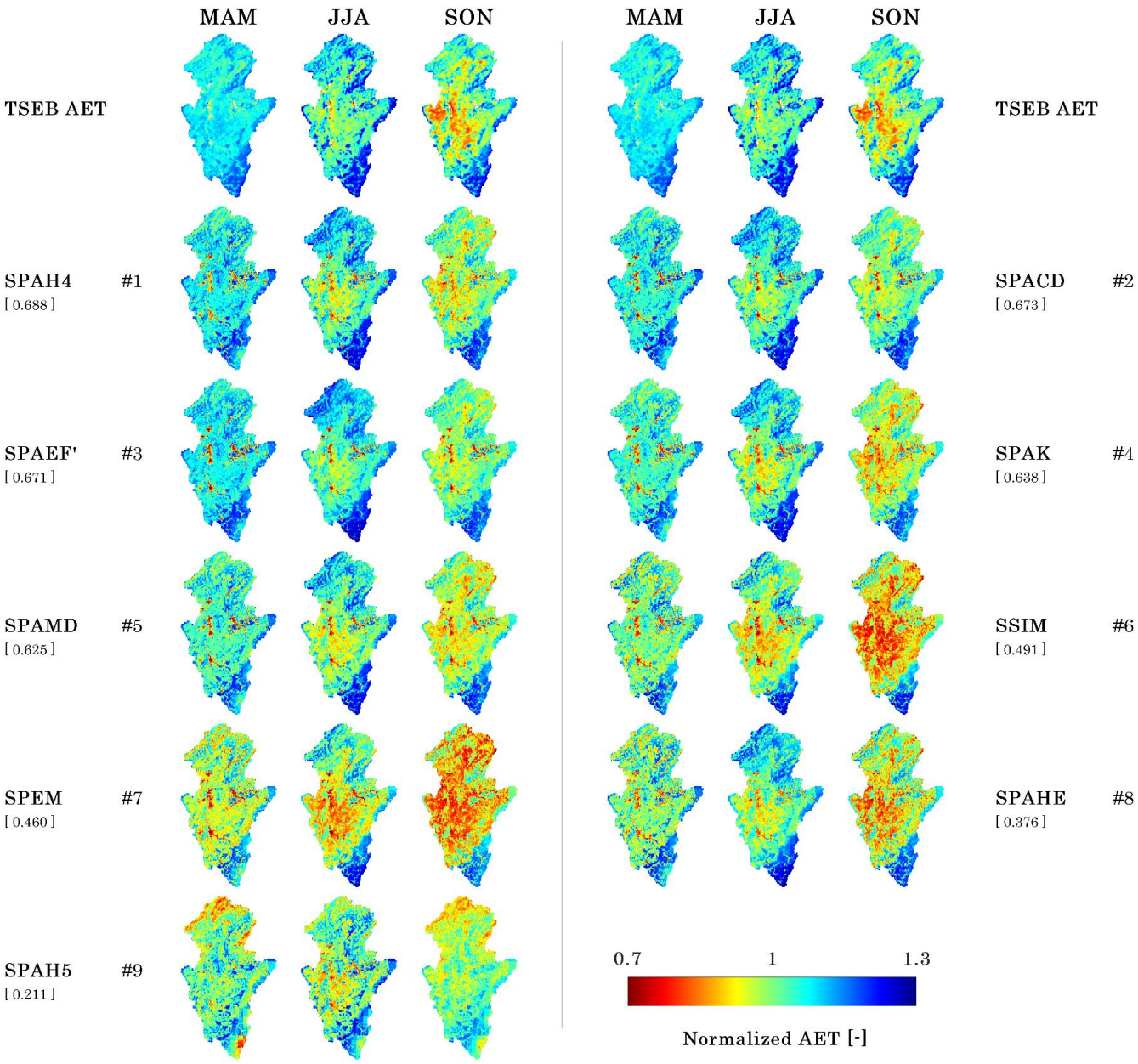
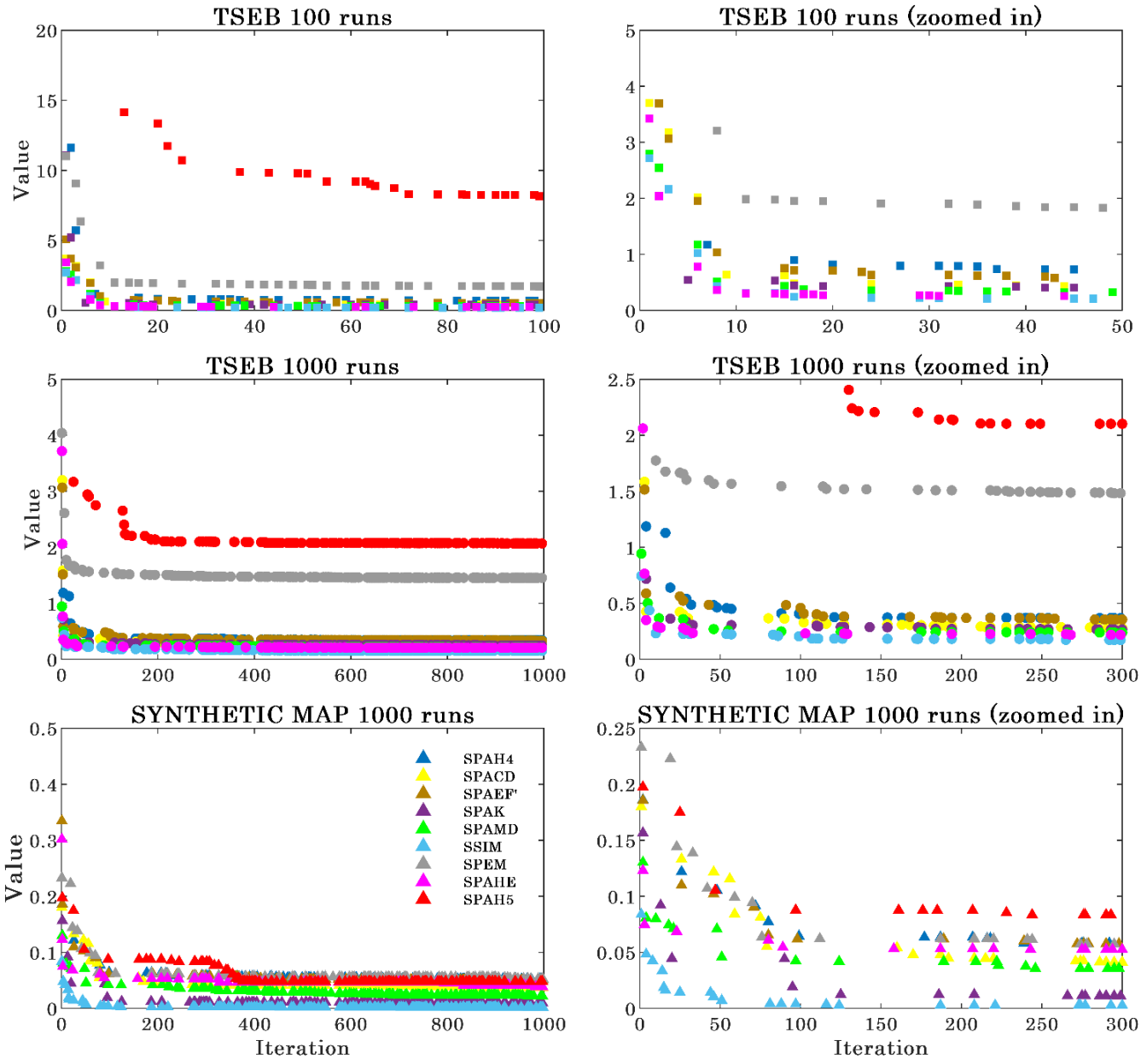



**Figure 3**. Long term average three-monthly TSEB reference maps versus mHM simulated maps using best-balanced Pareto solution parameter set from 1000 run case.

The entire calibration development process, the model improvements from beginning to end and the optimum points are depicted with scatter diagrams in Figure 4. It shows the relationship between the value and iteration based on the ParaPADDS search algorithm, more specifically, the objective function value achieved for each iteration step of the calibration process. While the OF results in Table 3 are obtained at the end of the iteration step sequence, some consistent metrics may reach this best value earlier. SPAH4 reached its best OF value at 0.70 and 0.35 in approximately quarter steps for 100 and 1000 runs, respectively. Similarly, SPACD, SPAEF' and SPAMD are also fast-improving metrics. Since the synthetic case was based on a virtually generated daily map, it took longer for the metrics to find the points where their improvement became linear, nearly a third. It is surprising to see that SPAH5 and SPEM are consistent early maturing metrics despite their poor spatial performance.

485




486

487

488 **Figure 4.** Scatter plots of the calibration processes, the OF value-iteration relationship of the
489 PDSS search algorithm. First and second column sub-plots are the same figures except for
490 different extent.

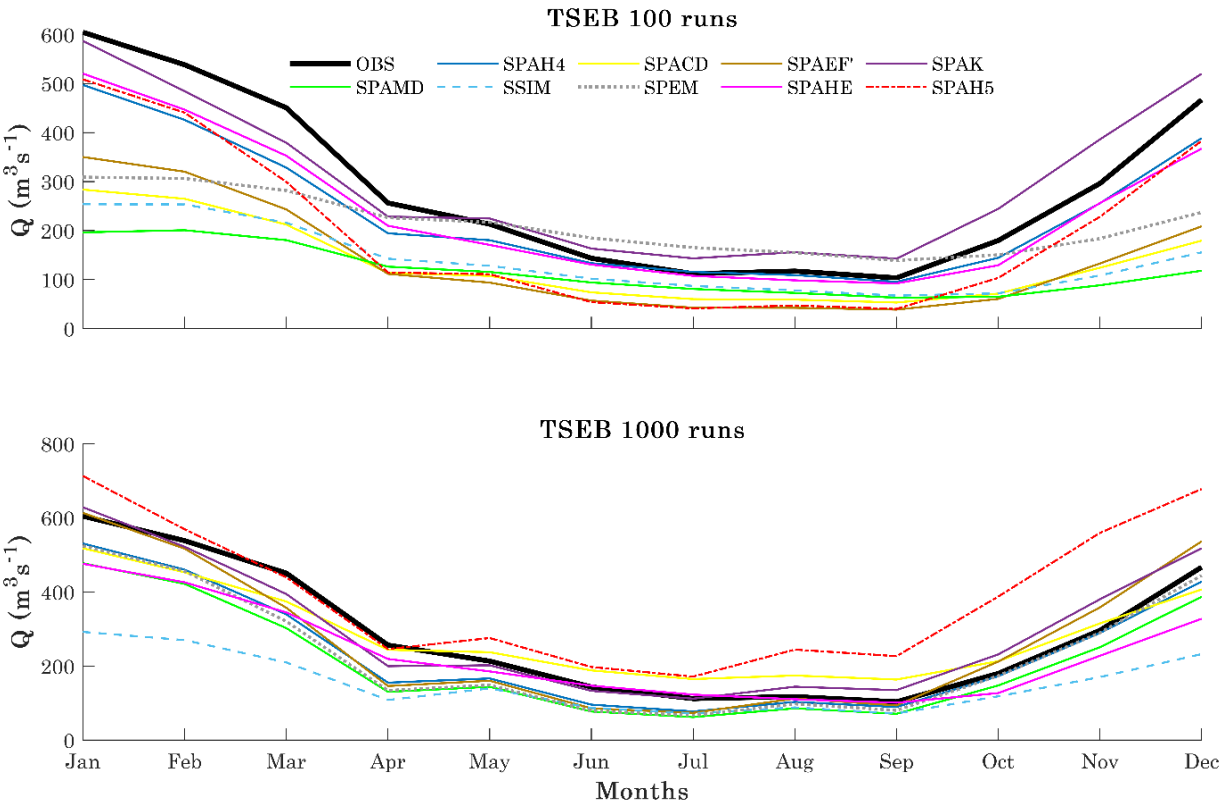



**Figure 5.** Monthly average hydrograph of all years in the calibration period (2002–2014) to demonstrate the flow simulation performances of nine different metrics.

Figure 5 compares in-situ observed hydrographs and simulated hydrographs constrained by metrics. SPAH4 and SPAK performed better in each case, predicting the most similar discharges to the observed Cochem outflows. Otherwise, the SPAHE metric standout for the 100 runs and the SPACD metric for the 1000 runs, as pointed out by the KGE column in Table 3. The simulations show better hydrograph fitting during the growing season, especially during the summer months, also the hydrograph line breakpoints, peaks and valleys are coherently followed. Thus, the overall trend and characteristics of the streamflow were successfully analyzed and represented. Also, a positive correlation was found between increasing iteration and hydrograph fit. As the number of iterations increases, the hydrograph lines become closer to the observed lines and the overall consolidation of the hydrographs provides better results. The narrow range of hydrographs in Figure 5 shows that the developed new metrics can be used not only for the spatial pattern performance simulating the AET but also for the temporal streamflow performance simulating the discharges.

Overall, the results indicate that the newly developed SPAH4 and SPACD are the best performing metrics for all calibration scenarios, particularly in the non-synthetic TSEB cases. The competitive performance of the SPAMD metric that follows them should not be ignored. Briefly, the four-component spatial performance metric SPAH4 stands out especially with its versatile evaluation and robust performance, indicated with bold text in Table 3. Although the modeler can use the SPAH4 and SPACD metrics in the long and short runs, respectively, both offer close values for the decision makers. We can see that the only negative output is experienced in the TSEB 100 runs i.e. SPAH5 It should not be overlooked that SPAH5 is a prominent metric for synthetic scenarios. Interestingly, there is a significant positive correlation between the KGE and the metrics containing the kurtosis statistic.

## 6 Discussion

This study sets out to assess the importance and comparison of spatial metrics in distributed model calibration. Previous studies have noted that spatial metrics are closer to the reference model than time series metrics in model optimizations (Demirel et al., 2018). One of the first objectives of the study is to select the appropriate spatial performance metric that plays an active role in simulating inadequate spatial AET models similar to satellite-based reference models. SPAEF has been the inspiration for this study with its innovations in spatial model parameterization and spatial performance metric selection. These innovations have raised new questions in the pattern comparison used in model optimization. Numerous imperfect models are produced during these optimizations, due to limitations in the chosen objective function. To overcome these limitations and to obtain a more physically meaningful and empowered metric, we have developed new metrics that include statistical and analytical approaches. Thanks to this meta-analysis, while suggesting the most successful metric for users, different objective functions that can be used for various purposes can also be seen as an opportunity. While searching for new solutions for a more robust spatial performance metric, we derived metrics that emphasize spatiality in a more comprehensive way by increasing the number of components of SPAEF and changing the content of the components. For the three cases, significant findings that are both different from each other and support each other have been identified. The TSEB 100 and 1000 run cases in model calibration served the purpose of evaluating metric performances in short and long runs, thus providing a flexible and versatile assessment that allows the progress of the model calibration performed by the metrics to be monitored and the decision maker to choose metrics according to their preferences.

TSEB 100 runs, which we tested by focusing on the performance of spatial metrics in short runs, SPACD and SPAMD demonstrated better results on the SPAEF basis compared to other metrics. Notably, SPAK and SPAH4 including the kurtosis coefficient ratio component, yielded the best KGE values even at iterations close to the beginning. TSEB 1000 runs which we tested by focusing on its performance in long runs, resulted in more decisive outcomes with no negative values for any criteria. SPAH4 emerged as the top-performing metric in this case, followed by SPACD. The competition between these metrics was notable. In the uncertainty analysis, SPAH4 has an acceptable sampling error although it has the extra component. (Table A1). Like the TSEB 100 runs, SPAK and SPAH4 exhibited the highest KGE values. This indicated consistency was strong evidence for important findings and suggests that the descriptive statistical kurtosis ratio component has a considerable positive effect on the discharge simulation. Due to the tendency of the SPAH4 metric including kurtosis for flow prediction, it worked as a metric that focused on both spatial and flow performance, although the analysis was performed with a single spatial performance-oriented objective function. It sheds light on the analysis in detecting the presence of outliers potential also differences in the tail and crests, controlling data integrity, understanding data distribution, reliability of the statistical analysis and improving the metric performance from a statistical perspective. Thus, by investigating and questioning the effect of outliers on spatial performance, the harmony and differences between them are also included in the model. Now that these outliers are introduced to the model, the histogram intercept component is also supported, the margin of error is reduced and a more exact match is made.

In the synthetic scenario, the metric SPAH5 which incorporates skewness characteristics, yielded the best SPAEF value. SPACD and SPAH4 also demonstrated successful outcomes in this scenario. The kurtosis information we use in the SPAH4 metric expresses how often outliers occur, while the skewness information we use as the fifth component in the SPAH5 metric gives information about the direction of the outliers. Our purpose in including the

565 skewness component is to question the likelihood of events in the probability distribution, and
566 especially to consider extreme distribution. Various datasets have different characteristics,
567 since the differences specific to this dataset represent important concepts in the calibration
568 model, many principles are referred to using the skewness information, from the algorithm of
569 the model to the physics-based hydrology information. Thus, we enabled a more
570 comprehensive and more specific analysis for models consisting of diverse data. Our finding
571 of the importance of these statistical measures in understanding the data is supported by the
572 study by Cain et al., processing skewness and kurtosis information on distributions collected
573 from the authors of the published articles (Cain et al., 2017). In addition, it is possible to derive
574 a positive interpretation from a negative finding in meta-analyses as in this study. Since the
575 only difference between the metrics with the best and the worst performance in TSEB runs,
576 namely SPAH4 and SPAH5, is the skewness ratio component, it can be concluded that
577 skewness is a component that negatively affects the spatial metrics used in pattern comparison.
578 It should be noted though that skewness information is an outstanding component for synthetic
579 cases.

580 In TSEB 100 runs scenario, the spatial performance tussle results of the metrics show that the
581 newly proposed metric i.e. SPACD outperforms the conventional three-component metric
582 SPAEF (5.76% better) on the other hand 11.11% better than SSIM and 1.66% better than
583 SPEM. In TSEB 1000 runs results demonstrate that the newly developed four-component
584 metric i.e. SPAtial Hybrid 4 (SPAH4) slightly outperform SPAEF (2.62% better). However,
585 SPAH4 significantly outperforms the other existing metrics i.e. 40.22% better than SSIM and
586 49.53% better than SPEM.

587 **7 Conclusion**

588 In this study, we thoroughly assessed common existing metrics and new spatial pattern-oriented
589 performance metrics that we developed based on SPAEF. For the consistency and reliability
590 of the results, the Mosel Basin with high data quality was selected and the physics-based fully
591 distributed mHM model was established for this basin. In these three different scenarios, we
592 performed analyses with various (low-high) iterations for actual evapotranspiration maps
593 (TSEB AET) and synthetic maps. The most popular metrics (SPAEF, SSIM and SPEM) were
594 compared with new metrics (SPAH4, SPACD, etc.) to measure the convergence of the mHM
595 model to long-term monthly AET maps observed during parameter calibration. The usage of
596 this synthetic scenario is important to ensure the reproducibility of the experiments and to give
597 us full control over the calibration process. Based on our findings we can draw the following
598 conclusions.

599 - The inclusion of kurtosis ratio coefficient in the spatial pattern-oriented metrics demonstrates
600 that metric performance is improved, so it has a positive impact on the spatially objective
601 functions. Also shows a positive effect on streamflow prediction, it successfully calibrates the
602 KGE metric even in very short runs. Furthermore, while using the skewness ratio coefficient
603 gave unsuccessful results for TSEB AET maps, the kurtosis information of the distribution was
604 more prominent in the pattern performance of the models. However, the SPAH5 performs the
605 best among the close results and is presented as a strong hypothesis for the synthetic cases.

606 - The metric with the best performance in the short runs was SPACD, which normalizes the
607 distribution according to density. The excellent consistency between histograms, which is the
608 main component of the Earth mover's distance metric, has a positive effect on making this
609 metric a sharp metric with little tolerance, making SPAMD the second-best metric.

- The best-performing metric on long runs was SPAH4, a four-component spatial performance metric that includes the kurtosis of the distribution. It was followed by the SPACD metric, which proved its consistent performance. Thus, the decision maker is presented with a flexible and wide working area.

- Considering all the experimental results, the most successful and robust metric in all three scenarios is our newly developed spatial pattern-oriented SPAH4, which outperforms the existing metrics in the literature by up to fifty per cent.

In future studies, it would significantly enhance the depth and quality of the analysis to increase the number of iterations. In fact, convergence in hydrological models is closely related to the number of parameters and the freedom of the appropriate iteration chosen. Future work may benefit from exploring untested statistical terms to add a new perspective. We expect that these newly developed metrics, especially SPAH4, will be used not only in hydrology but also in other fields including remote sensing, image processing and object detection.

**Appendix A:** Results of the jackknife and bootstrap based sampling uncertainty analysis. Clark et al (2021) showed that the two most popular metrics in hydrology, i.e. NSE and KGE, are vulnerable to sampling uncertainty since the differences between observed and simulated streamflow values at random time steps in time series which can have significant effects on the results (Knoben & Spieler, 2022). From this study, we are inspired to assess the sampling uncertainty in ten metrics using the gumboot R package (Clark et al., 2021) which uses a jackknife-after-bootstrap method of Efron (1992) to estimate standard errors (SEJaB) shown in Table A1.

**Table A1.** Sampling uncertainty of the metrics i.e. ranked based on the seJab column.

| GOF_stat | seJack | seBoot | p05 | p50 | p95 | score | biasJack | biasBoot | seJab |
|---|---|---|---|---|---|---|---|---|---|
| **SSIM** | 0.0103 | 0.0099 | 0.6144 | 0.6311 | 0.6457 | 0.6307 | -0.0002 | 0.0000 | 0.0091 |
| **SPAHE** | 0.0568 | 0.0119 | 0.7717 | 0.7917 | 0.8107 | 0.7783 | 0.1496 | 0.0131 | 0.0112 |
| **SPAMD** | 0.0114 | 0.0108 | 0.6785 | 0.6972 | 0.7137 | 0.6966 | 0.0006 | 0.0001 | 0.0115 |
| **SPEM** | 0.0180 | 0.0175 | 0.2739 | 0.3041 | 0.3309 | 0.3034 | -0.0006 | -0.0003 | 0.0146 |
| **SPAEF** | 0.0133 | 0.0128 | 0.6489 | 0.6711 | 0.6917 | 0.6727 | 0.0017 | -0.0021 | 0.0148 |
| **SPAEF'** | 0.0133 | 0.0127 | 0.6489 | 0.6711 | 0.6917 | 0.6727 | 0.0017 | -0.0021 | 0.0152 |
| **SPAK** | 0.0302 | 0.0288 | 0.5719 | 0.6226 | 0.6661 | 0.6207 | -0.0004 | 0.0007 | 0.0295 |
| **SPAH4** | 0.0302 | 0.0298 | 0.5484 | 0.5999 | 0.6459 | 0.6000 | 0.0011 | -0.0012 | 0.0313 |
| **SPACD** | 0.0234 | 0.0248 | 0.6056 | 0.6571 | 0.6851 | 0.6670 | -0.0219 | -0.0142 | 0.0603 |
| **SPAH5** | 0.1685 | 0.2077 | -0.3594 | 0.0373 | 0.2636 | 0.0427 | -0.0388 | -0.0401 | 0.3382 |

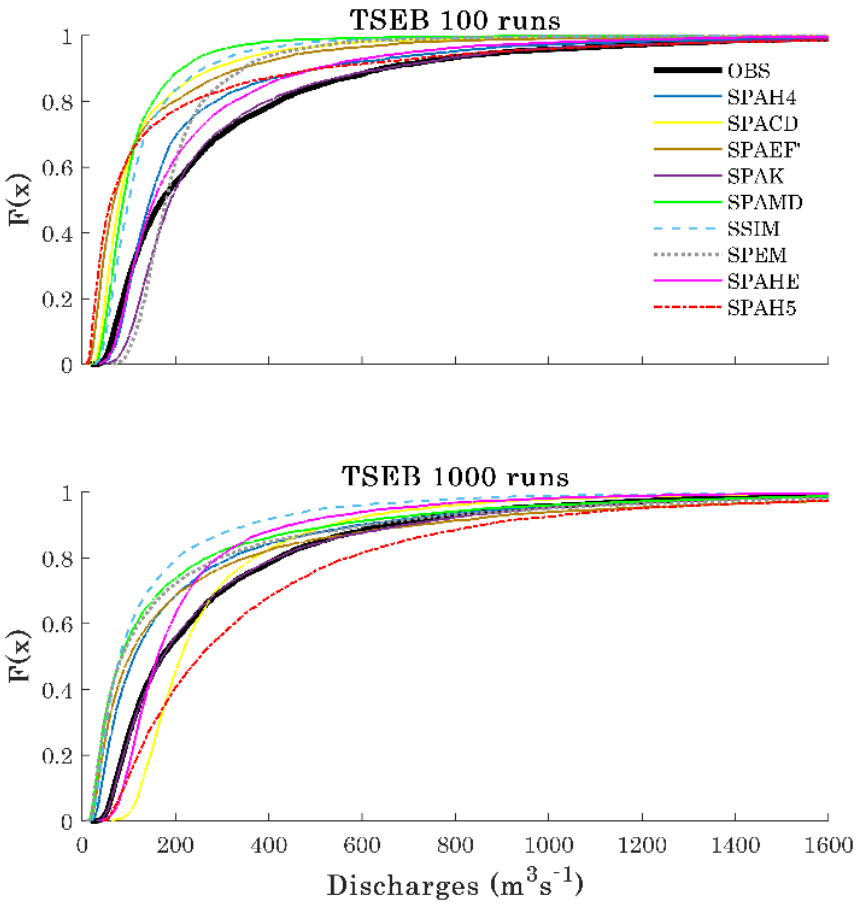



**Figure A1.** eCDF plot of daily discharge for all years in the calibration period (2002-2014) to visualize the distribution of the data and identify statistical patterns.

Figure A1 visualizes the empirical cumulative distribution function (eCDF) plot for the observed and simulated data, which shows how the probability of a given discharge value occurring varies over the range of discharge values. In this context, the percentage of observed discharges less than nearly 500 is 80% and less than 200 is 50% for both the TSEB 100 and 1000 runs. Furthermore, the slope of the curve at any point represents the density function of the discharge values at that point, and the intervals where the curve steepens contain values close to the mean value. Hence, it can be concluded that the overall average discharge value of the steepening intervals of the flow data resulting from the simulation of the metrics is roughly 300 $m^3/s$. The mean observed outflow of Cochem station is around 315 $m^3/s$ supports this outcome. In both cases, SPAK and SPAH4 illustrated a high level of matching in terms of the fit of the curves generated by the observed data (OBS) and the metrics, with the least difference between the distributions.

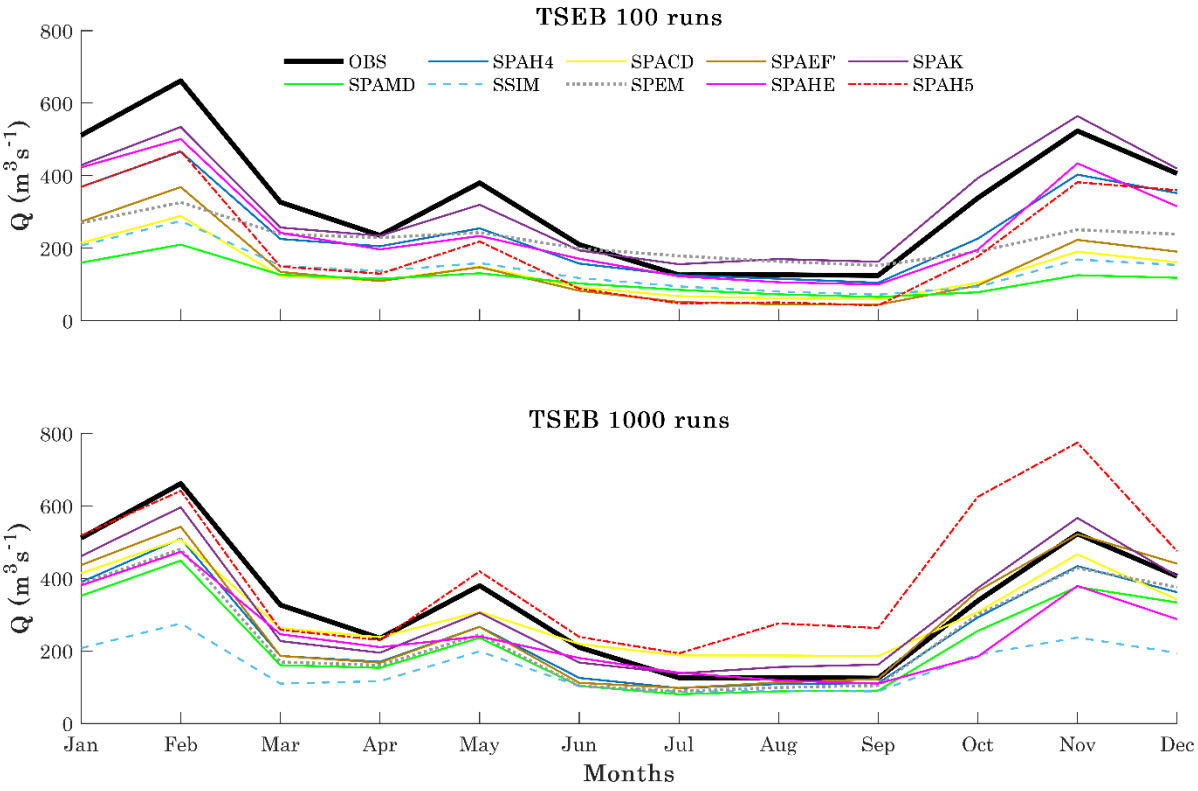



**Figure A2.** Monthly average hydrograph of the last two years in the calibration period (2013–2014)

## 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 model calibration software Ostrich is available from https://github.com/usbr/ostrich. The simulation scripts and results of the mHM model simulations are publicly available at https://doi.org/10.5281/zenodo.8059198. The source code to quantify the sampling uncertainty in performance metrics (the "gumboot" package) is available at https://github.com/CH-Earth/gumboot. The scripts and results of the gumboot-based sampling uncertainty analysis is available at https://doi.org/10.5281/zenodo.8058659

**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 NASA program i.e. NNH22ZDA001N-RRNES: A.24 Rapid Response and Novel Research in Earth Science under the grant number 22-RRNES22-0010 and by the Scientific Research Projects Department of Istanbul Technical University (ITU-BAP) under grant number MDA-2022-43762 and by 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


Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., & Chung, E.-S. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences*, *23*(11), 4803–4824. https://doi.org/10.5194/hess-23-4803-2019

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *Crop evapotranspiration - Guidelines for computing crop water requirements.* FAO Irrigation and drainage paper 56. Retrieved from http://www.fao.org/docrep/x0490e/x0490e00.htm (accessed at 16/02/2018)

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., et al. (2021). Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, *3*(6). https://doi.org/10.1148/ryai.2021200267

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

Avcuoğlu, M. B., & Demirel, M. C. (2022). Hidrolojik Model Kalibrasyonunda Uydu Tabanlı Aylık Buharlaşma ve LAI Verilerinin Kullanılması. *Teknik Dergi*, *33*(6), 13013–13035. https://doi.org/10.18400/tekderg.1067466

Becker, R., Koppa, A., Schulz, S., Usman, M., aus der Beek, T., & Schüth, C. (2019). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived ET data. *Journal of Hydrology*, *577*(10), 123944. https://doi.org/10.1016/j.jhydrol.2019.123944

Beven, K. (2023). Benchmarking hydrological models for an uncertain future. *Hydrological Processes*, *37*(5). https://doi.org/10.1002/hyp.14882

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., et al. (2017). Looking beyond general metrics for model comparison – lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

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

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

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

Danapour, M., Fienen, M. N., Højberg, A. L., Jensen, K. H., & Stisen, S. (2021). Multi-Constrained Catchment Scale Optimization of Groundwater Abstraction Using Linear Programming. *Groundwater*, *59*(4), 503–516. https://doi.org/10.1111/gwat.13083

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., Booij, M. J., & Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resources Research*, *49*(7), 4035–4053. https://doi.org/10.1002/wrcr.20294

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2015). The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. *Hydrology and Earth System Sciences*, *19*(1), 275–291. https://doi.org/10.5194/hess-19-275-2015

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

Dougherty, E., Sherman, E., & Rasmussen, K. L. (2020). Future Changes in the Hydrologic Cycle Associated with Flood-Producing Storms in California. *Journal of Hydrometeorology*, *21*(11), 2607–2621. https://doi.org/10.1175/JHM-D-20-0067.1

Efford, N. (2000). Digital Image Processing: A Practical Introduction Using JavaTM. Pearson Education. Slate.

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

Fréchet, M. M. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti Del Circolo Matematico Di Palermo*, *22*(1), 1–72. https://doi.org/10.1007/BF03018603

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator:L 2 theory. *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, *57*(4), 453–476. https://doi.org/10.1007/BF01025868

Gaur, S., Singh, B., Bandyopadhyay, A., Stisen, S., & Singh, R. (2022). Spatial pattern-based performance evaluation and uncertainty analysis of a distributed hydrological model. *Hydrological Processes*, *36*(5). https://doi.org/10.1002/hyp.14586

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

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

Hausdorff, F. (1914). Bemerkung über den Inhalt von Punktmengen. *Mathematische Annalen*, *75*(3), 428–433.

Hossain, M. K., & Meng, Q. (2020). A thematic mapping method to assess and analyze potential urban hazards and risks caused by flooding. *Computers, Environment and Urban Systems*, *79*, 101417. https://doi.org/10.1016/j.compenvurbsys.2019.101417

Immerzeel, W. W., & Droogers, P. (2008). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of Hydrology*, *349*(3–4), 411–424.

https://doi.org/10.1016/j.jhydrol.2007.11.017

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

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences Discussions*, (July), 1–7. https://doi.org/10.5194/hess-2019-327

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

López, P., Sutanudjaja, E. H., Schellekens, J., Sterk, G., & Bierkens, M. F. P. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, *21*(6), 3125–3144. https://doi.org/10.5194/hess-21-3125-2017

Martinez-Villalobos, C., Neelin, J. D., & Pendergrass, A. G. (2022). Metrics for Evaluating CMIP6 Representation of Daily Precipitation Probability Distributions. *Journal of Climate*, *35*(17), 5719–5743. https://doi.org/10.1175/JCLI-D-21-0617.1

Matott, L. Shawn. (2004). *OSTRICH: an Optimization Software Tool, Documentation and User's Guide, Version 17.12.19.* Retrieved from https://github.com/usbr/ostrich

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

Monteith, J. L. (1965). EVAPORATION AND ENVIRONMENT. *Symposia of the Society for Experimental Biology*, *19*, 205–234.

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

Nilsson, J., & Akenine-Möller, T. (2020). Understanding SSIM. Retrieved from https://arxiv.org/abs/2006.13846

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

Onyutha, C. (2022). A hydrological model skill score and revised R-squared. *Hydrology Research*, *53*(1), 51–64. https://doi.org/10.2166/nh.2021.071

Pearson, K. (1905). The problem of the random walk. *Nature*, *72*(1865), 294.

Priestley, C. H. B., & Taylor, R. J. (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, *100*(2), 81–92.

https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2

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

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

Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (pp. 59–66). Narosa Publishing House. https://doi.org/10.1109/ICCV.1998.710701

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., Brenner, J., Craven, J., Cuntz, M., Dalmasso, G., Demirel, C. M., et al. (2021, July 21). The mesoscale Hydrologic Model - mHM v5.11.2. https://doi.org/10.5281/ZENODO.5119952

Schneider, R., Henriksen, H. J., & Stisen, S. (2022). A robust objective function for calibration of groundwater models in light of deficiencies of model structure and observations. *Journal of Hydrology*, *613*, 128339. https://doi.org/10.1016/j.jhydrol.2022.128339

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610. https://doi.org/10.1093/biomet/66.3.605

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

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., & Jensen, K. H. (2018). Moving beyond run-off calibration-Multivariable optimization of a surface-subsurface-atmosphere model. *Hydrological Processes*, *32*(17), 2654–2668. https://doi.org/10.1002/hyp.13177

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, *21*(153), 65–66. https://doi.org/10.1080/01621459.1926.10502161

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32. https://doi.org/10.1007/BF00130487

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

Thoya, P., Maina, J., Möllmann, C., & Schiele, K. S. (2021). AIS and VMS Ensemble Can Address Data Gaps on Fisheries for Marine Spatial Planning. *Sustainability*, *13*(7), 3769. https://doi.org/10.3390/su13073769

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for

847 computationally efficient watershed model calibration. *Water Resources Research*, *43*(1).
848 https://doi.org/10.1029/2005WR004723

849 Wakigari, S. A., & Leconte, R. (2023). Assessing the Potential of Combined SMAP and In-
850 Situ Soil Moisture for Improving Streamflow Forecast. *Hydrology*, *10*(2), 31.
851 https://doi.org/10.3390/hydrology10020031

852 Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment:
853 From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*,
854 *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

855 Wiederholt, R., Paudel, R., Khare, Y., Davis, S. E., Melodie Naja, G., Romañach, S., et al.
856 (2019). A multi-indicator spatial similarity approach for evaluating ecological restoration
857 scenarios. *Landscape Ecology*, *34*(11), 2557–2574. https://doi.org/10.1007/s10980-019-
858 00904-w

859 Yoo, S. B. M., Tu, J. C., Piantadosi, S. T., & Hayden, B. Y. (2020). The neural basis of
860 predictive pursuit. *Nature Neuroscience*, *23*(2), 252–259. https://doi.org/10.1038/s41593-
861 019-0561-6

862 Zink, M., Mai, J., Cuntz, M., & Samaniego, L. (2018). Conditioning a Hydrologic Model Using
863 Patterns of Remotely Sensed Land Surface Temperature. *Water Resources Research*,
864 *54*(4), 2976–2998. https://doi.org/10.1002/2017WR021346

865