Global reference evapotranspiration clustering and its relation to the Köppen-Geiger climate classification

5 authors, including:

Nikolaos Malamos
Agricultural University of Athens
76 PUBLICATIONS   **589** CITATIONS

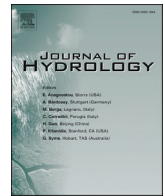Aristoteles Tegos
Hydrology
54 PUBLICATIONS   **793** CITATIONS

Demetris Koutsoyiannis
National Technical University of Athens
1,031 PUBLICATIONS   **19,823** CITATIONS

Global reference evapotranspiration clustering and its relation to the Köppen-Geiger climate classification

Nikolaos Malamos [a,b,*] , Aristotelis Tegos [c], George Bourantas [b], Christos Chalvantzis [b], Demetris Koutsoyiannis [c]

[a] Department of Natural Resources Development and Agricultural Engineering, Agricultural University of Athens, Iera Odos 75, GR-11855 Athens, Greece
[b] Department of Agriculture, University of Patras, Nea Ktiria Campus, GR-30200 Messolonghi, Greece
[c] Department of Water Resources and Environmental Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece

ABSTRACT

This study introduces a global clustering approach for reference evapotranspiration (ETo) based on spectral clustering techniques. By analyzing 3139 stations worldwide, the method segments the globe into 45 unique clusters that reflect shared ETo dynamics. This clustering technique not only enhances the accuracy of ETo estimation but also provides a simplified method for computing ETo in areas where meteorological data are limited. The spatial extent of these clusters allows for the application of pre-calculated parameters, streamlining ETo calculation and reducing the computational complexity typically involved. This approach draws an analogy to the Köppen-Geiger climate classification, aligning ETo clusters with well-established climate zones, thus revealing patterns between evapotranspiration and global climate behavior. Moreover, the study demonstrates that the clusters can be utilized to calibrate other ETo estimation methods, offering a tool for guiding regional model adjustments. The validation of this method, using 30 stations from diverse climates, showed strong correlations and acceptable performance metrics, indicating its applicability for hydrological and engineering tasks. In conclusion, this global ETo clustering provides a robust, accessible framework for water resource management, irrigation planning, and climate adaptation strategies, particularly in data-scarce regions.

1. Introduction

Evapotranspiration (ET) represents a historically critical component of the global water cycle, as it encapsulates the combined processes of water evaporation from land surfaces and transpiration from plants (Monteith, 1981; Penman, 1948; Ritchie, 1972). For agricultural, hydrological, and environmental applications, the accurate estimation of evapotranspiration, specifically reference evapotranspiration (ETo), is vital in assessing crop water needs, planning irrigation schedules, and modeling hydrological processes under varying climatic conditions. ETo represents the amount of evapotranspiration that would occur under a standard set of conditions, assuming a specific crop (such as grass) under well-watered conditions, providing a standardized measure across different climates and ecosystems (Allen et al., 1998; Hargreaves and Allen, 2003; Hargreaves and Samani, 1982; Tegos et al., 2015b). Its calculation is a foundation for understanding more specific evapotranspiration needs for various crops, especially under stressful conditions like droughts or heatwaves.

Traditionally, ETo has been calculated using, temperature, radiation-temperature, combination, and multivariate models, such as the Penman-Monteith equation (PM) (Allen et al., 1998; Doorenbos and Pruitt, 1977; McMahon et al., 2016), which integrates meteorological variables like temperature, solar radiation, humidity, and wind speed to provide a reference estimate. While the PM equation is widely used, it requires a substantial amount of meteorological data, which may not be consistently available in many regions around the world.

Current methodologies for ETo estimation, while numerous and varied (McMahon et al., 2016), often require calibration to tackle spatial specificity and sensitivity to local climatic variations. This limitation hampers the precision of water resource planning and agricultural irrigation scheduling, leading to inefficiencies and increased vulnerability to climatic anomalies (Pereira et al., 2020). To tackle this, machine learning techniques and clustering methods have emerged as powerful tools to enhance the precision of ETo estimates, particularly in areas where data availability is limited or where climate conditions vary significantly over space and time (Antonopoulos and Antonopoulos,

2017; Umutoni and Samadi, 2024). By leveraging these advanced techniques, researchers have been able to model ETo more effectively, even with incomplete data, and have discovered new patterns in evapotranspiration behavior across diverse landscapes (Ippolito et al., 2024).

For example, in Iran, Kousari et al. (2013) provided a large-scale survey of ETo trends, covering the period from 1960 to 2005. Their study involved the use of clustering techniques to identify spatial and temporal patterns in ETo across the country. By grouping regions with similar ETo trends, they were able to assess the impacts of climate variability on water resources and agricultural productivity. The study underscored the utility of clustering as a tool for regionalizing ETo estimates and highlighted the long-term changes in evapotranspiration trends, which are critical for understanding the impacts of climate change on water availability.

Meanwhile, Demertzi et al. (2016) conducted an in-depth study on the hydroclimatic dynamics of Greece by applying multi-parametric clustering to both monthly precipitation and ETo data. Their research involved the identification of distinct regions based on similarities in climatic conditions and the generation of clusters that reflect the spatial variability of ETo across the country. The study highlighted the effectiveness of clustering in capturing complex climatic interactions that influence evapotranspiration, especially in geographically diverse regions such as Greece, where mountainous terrain and coastal areas create significant local climate variations.

Similarly, Bellido-Jiménez et al. (2022) explored the potential of machine learning and clustering in Southern Spain, a region characterized by frequent droughts and water scarcity. Their study introduced a novel approach by applying a regional machine learning-based clustering method to improve ETo estimation, particularly in data-scarce regions. The authors demonstrated that this clustering approach outperformed traditional temperature-based models by grouping meteorological data into clusters with similar ETo behaviors, which improved the overall accuracy of the ETo estimates. The results of the study underscored the importance of leveraging regional data and clustering techniques in enhancing the performance of evapotranspiration models, especially in arid and semi-arid regions where accurate water management is essential.

In Brazil, where climatic variability is vast due to the country's size and geographical diversity, Ferreira et al. (2022) applied a clustering-based approach to improve ETo estimates across different regions. Brazil's climatic zones range from tropical rainforests to arid semi-deserts, which present unique challenges for accurate ETo modeling. Ferreira et al. used multi-task learning in combination with clustering to integrate meteorological data and ETo estimates, producing more accurate models for regions with limited meteorological information. This study highlighted the power of combining clustering methods with machine learning to overcome data limitations and ensure precise ETo estimation in diverse climatic regions.

Di Nunno and Granata, (2023) applied a similar clustering approach to predict future ETo trends in the island of Sicily, using CORDEX climate data. CORDEX (Coordinated Regional Climate Downscaling Experiment) provides high-resolution climate data for different regions, and its integration with machine learning techniques has allowed for more precise regional forecasts of evapotranspiration under changing climatic conditions. Di Nunno and Granata's work combined clustering and machine learning to project future ETo values and identify potential water stress scenarios in Sicily. By clustering areas with similar ETo trends, they were able to create targeted predictions, providing valuable insights for future agricultural planning and water resource management.

To our knowledge, a global-scale clustering analysis of monthly reference ETo has not yet been conducted. This study introduces a novel framework for understanding the spatial distribution of ETo rates and their intricate relationship with climatic and geographic factors. By segmenting the globe into distinct clusters based on parameters with

physical meaning, our approach reveals underlying spatial patterns and correlations, ensuring that the outcome captures regional environmental influences. These spatially coherent clusters serve as a foundation for refining and calibrating both new and existing ETo estimation methods, providing region-specific parameterizations that enhance model accuracy. Furthermore, spatial contiguity enhances the applicability of the results, facilitating their integration into hydrological models, developing more effective water resource management strategies, optimizing irrigation systems, and mitigating the impacts of climate variability on agriculture and ecosystems.

Apart from the distribution analysis by country, the produced clusters are analyzed in relation to the Köppen-Geiger climate classification, which categorizes global climates based on specific temperature and precipitation thresholds. While evapotranspiration is inherently influenced by climatic variables, limited past research has explored the connection between evapotranspiration patterns and climate classifications (Akhavan et al., 2018; Almorox et al., 2015; Pimentel et al., 2023). Through cross-tabulation and percentage distribution analyses, we acquired the Köppen-Geiger classes distribution within each ETo cluster, enabling a detailed exploration of the interplay between ETo patterns and global climate zones.

2. Materials and methods

2.1. The parametric model

The Parametric model (Tegos et al., 2015a,b;2013), provides evapotranspiration estimates through calibration based on given evapotranspiration data. The model's performance was satisfactory as the proposed framework provides consistent monthly evapotranspiration estimates at point and especially at regional scale. The calibration can be performed for either Potential Evapotranspiration (PET) or ETo, depending on the intended application. In this study, the model was calibrated using monthly ETo data, ensuring that the outcome aligns with standardized reference crop conditions.

The mathematical expression of the parametric model, which is applicable to temporal scales from daily to monthly, is the following:

$$\text{ETo} = \frac{a\,R_a + b}{1 - c\,T} \tag{1}$$

where ETo is the reference evapotranspiration in kg m^{-2}, which is equivalent to mm, R_a (kJ m^{-2}) is the extraterrestrial radiation, T (°C) is the mean air temperature, and a (kg kJ^{-1}), b (kg m^{-2}) and c (°C^{-1}) are model parameters that should be inferred through calibration, against evapotranspiration data, either modelled or measured. We remark that from a macroscopic point-of-view, the above parameterization has some physical correspondence to the Penman-Montheith (PM) equation, since the product $a\,R_a$ represents the overall energy term (i.e., incoming minus outgoing solar radiation), parameter b represents the missing aerodynamic term. Finally, the expression $(1 - c\,T)$ approximates the denominator term of the Penman-Montheith formula, i.e. $(1 + \gamma'/\Delta)$, given the fact that γ' is a function of the surface and aerodynamic resistance and Δ is the slope of the vapour pressure curve, which is a function of temperature (Tegos et al., 2015b).

Tegos et al. (2017), in the context of producing global maps of optimized model parameters, argued that a more parsimonious version of the Parametric Model, tackles uncertainties due to "blind" calibration approaches or overfitting, when the actual ETo data is little informative to support the inference of the three parameters, e.g. due to limited length of associated meteorological data:

$$\text{ETo} = \frac{a'R_a}{1 - c'(T_{\min} + T_{\max})/2} \tag{2}$$

Equation (2) contains two instead of three parameters (parameter a' in the numerator and parameter c' in the denominator) also considering

the minimum and maximum temperature, instead of the mean daily one.

2.2. ETo clustering using a' and c' parameters

The Parametric approach allows for mapping the spatial distribution of the optimized model parameters a' and c', instead of its response, i.e. ETo (Malamos et al., 2017). This is a major advantage, since it allows implementing eq. (2) wherever in the globe, using interpolated values of the point (i.e., locally calibrated) parameters. Since the two parameters are negatively correlated (Tegos et al., 2017), they reflect the correlation of the associated meteorological variables of the parametric formula (extraterrestrial radiation, in the numerator, and temperature, in the denominator).

While clustering directly based on ETo measurements might seem intuitive, it potentially overlooks the complex interplay of climatic factors that influence ETo. The chosen methodology, by leveraging the Parametric method's parameters and geographic locations, provides a comprehensive view of the climatic and environmental patterns affecting ETo, yielding clusters that are both climatically meaningful and geographically coherent. This approach ensures that the resulting clusters are not only scientifically robust but also practically applicable for addressing global water resources and agricultural challenges.

The choice to produce clusters based on the two parameters (a' and c') from the Parametric method using weights that combine geographic location of each station along with the corresponding values of the parameters, versus directly using ETo measurements for clustering, can be justified on several grounds.

First of all, this approach reflects the underlying climatic processes, since the parameters a' and c' in the Parametric method, as already presented, encapsulate critical climatic influences on ETo, including temperature, solar radiation, wind speed, and humidity. Clustering based on these parameters rather than raw ETo values allows for groups to be formed based on underlying climatic processes that drive ETo, offering a more nuanced understanding of evapotranspiration dynamics across different regions. Incorporating these parameters, even as a minor component, ensures that clusters have a degree of climatic coherence, which is crucial for practical applications like agricultural planning and water resource management. This approach mitigates the risk of creating clusters that are geographically similar but climatically dispersed.

Furthermore, the direct use of ETo measurements may be influenced by local anomalies or short-term fluctuations that do not accurately reflect the broader climatic trends. The Parametric method's parameters are conceived to model ETo across a wide range of conditions. Using these parameters for clustering is inherently robust to the diverse climates encountered globally. By focusing on a' and c' parameters, clusters represent stable, long-term climatic similarities among stations, reducing the impact of temporary variations and ensuring that clusters capture consistent environmental characteristics.

2.3. ETo data

The FAO CLIMWAT 2.0 database is a joint initiative by the Water Development and Management Unit and the Climate Change and Bioenergy Unit of Food and Agriculture Organization of the United Nations (FAO, 1993), which provides average monthly climatic data at 4300 stations, well-distributed worldwide. These data include monthly mean values of mean daily maximum and minimum temperature (°C), daily relative humidity (%), wind speed (km day^{-1}), daily sunshine duration (h), daily solar radiation (MJ/m^2), monthly rainfall, gross and effective (mm), as well as mean monthly ETo estimations through the Penman-Monteith formula.

Tegos et al. (2017), used the FAO CLIMWAT 2.0 database to calibrate the parametric expression (eq. (2)), thus providing local estimations of parameters a' and c' at all station sites, using the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) as the objective function to

maximize against parameters a' and c'. The model demonstrated strong predictive performance across Eurasia, North America, and much of Oceania, as confirmed by multiple evaluation metrics. However, its performance was weaker in equatorial regions, including South America, Africa, and the Indian and Indonesian Peninsula. This lower accuracy, as indicated by the NSE criterion, is likely due to the model's omission of relative humidity and wind speed, two crucial factors influencing net incoming solar radiation and evaporation demand, which are key drivers of the evapotranspiration process in tropical climates (Tegos et al., 2017). Thus, to enhance the validity of our research outcomes, we limited our selection to stations that demonstrated NSE values above 0.75, resulting in a refined dataset of 3139 stations worldwide.

2.4. Spectral clustering

Clustering methods like k-means, k-medians or k-medoids are designed to discover convex clusters in multidimensional data (Han and Kamber, 2012). In the context of the present study, we employed all the mentioned clustering methods to create clusters of reference evapotranspiration across the globe using the parameters of the Parametric ETo model, but according to our findings they failed to create good separations. Our analysis showed that spectral clustering not only achieved higher consistency in cluster quality but also provided meaningful clusters of the a' and c' parameters data.

Spectral clustering uses the eigenvalues of the affinity matrix to project data into a lower-dimensional space, where clusters can be more easily identified (Jia et al., 2014). The affinity matrix is a square matrix where each entry (i, j) represents the degree of similarity or "affinity" between two data points i and j. The values in the affinity matrix are typically based on a similarity function, such as a Gaussian (radial basis function) kernel, that quantifies how close or similar the points are in a given space. The matrix is symmetric, with higher values indicating stronger similarities and lower values indicating less similarity or distance between points.

This approach ensures that the true relationships between different geographic zones are better maintained and reflected in the resulting clusters, particularly in terms of handling complex data and capturing the intrinsic similarities within the dataset. The dimensionality of the new space is set to the desired number of clusters. This setting expects that each new dimension should be able to manifest a cluster (Han and Kamber, 2012).

A brief description of the spectral clustering follows.

Given a set of objects $o_1, ..., o_n$, the distance between each pair of objects, dist(o_i, o_j) with $1 \leq i, j \leq n$, and the desired number k of clusters, the spectral clustering workflow includes the following steps (Von Luxburg, 2007):

1. Using the distance measure, calculate an affinity matrix, **W**, using pairwise similarities between data points to form it. This is achieved by computing a k-nearest neighbors (k-NN) graph.
2. Via the affinity matrix **W**, derive the normalized graph Laplacian **L**, which encapsulates the data's connectivity as:

$$L = I - D^{-1/2} \, W \, D^{-1/2} \qquad (3)$$

where **I** is the identity matrix and **D** is the degree matrix. The degree matrix **D** is a diagonal matrix where each entry D_{ii} represents the sum of the affinities between point i and every other point j:

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \qquad (4)$$

3. Find the k leading eigenvectors of **L**. By performing eigen decomposition on the Laplacian matrix **L**, select the eigenvectors corresponding to the smallest non-zero eigenvalues, as these capture the

most significant structural information of the data. The number of selected eigenvectors typically corresponds to the desired number of clusters k.

4. Using the k leading eigenvectors, project the original data into the new lower-dimensional space defined by those eigenvectors and run a clustering algorithm such as k-means to find k clusters.

2.5. Cluster size analysis

For the needs of the present study, following spectral clustering we applied the k-means method to acquire the size, i.e. the number, of the clusters of the ETo dataset, using a centroid-based partitioning technique implementing spatial weights based on multi-attribute distances, i.e. parameters' values and geometric distance between the stations and the cluster's center point. The centroid is defined as the mean value of the a' and c' parameters from the stations assigned to the cluster (Han and Kamber, 2012), while in terms of geographical distance, the centroid is defined as the geometric difference between a station and the center of the cluster, determined as the great circle distance (Earle, 2005; Tseng and Lee, 2007) between the two points. The quality of clustering was measured by the within-cluster variation, which was the sum of squared error (SSE) between all stations in a cluster and its centroid, computed as the multi-attribute distance of a' and c' parameters and the geometric centroid. Thus, SSE acts as a global measure of error, as the number of clusters increases, the SSE decreases because clusters are, by definition, smaller:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{s=1}^{h} dist(p_s, c_i)^2 \tag{5}$$

where $dist(p_s, c_i)$ is the multi-attribute distance of each station s and the centroid of cluster c_i ($1 \leq i \leq k$) that includes h stations, for k clusters in total.

All variables, i.e. a', c', and geometric distance, were normalized to a common scale by subtracting their respective mean and dividing by the mean absolute deviation (MAD) values (Kaufman and Rousseeuw, 2005). Thus, each variable is transformed to have a mean of 0 and a dispersion based on MAD, ensuring that all variables contribute equitably to the clustering process:

$$X_{\text{norm}} = \frac{X - \overline{X}}{\text{MAD}(X)}, \text{MAD}(X) = \frac{1}{n} \sum_{i=1}^{n} |X_i - \overline{X}| \tag{6}$$

where X_{norm} is the normalized value, X represents each variable, i.e. a', c', or geometric distance, and $\overline{X}$ is the mean of the corresponding variable.

The determination of the appropriate cluster size was based in the comparison of the SSE for a different numbers of cluster solutions, using the elbow method (Han and Kamber, 2012). Elbow method incorporates a plot of the SSE against the series of sequential cluster levels and provides a graphical way to choose an appropriate cluster size, such that the intra-cluster similarity is maximized, and the between-cluster similarity is minimized (Anselin, 2024). That is, an appropriate cluster solution could be defined as the solution at which the reduction in SSE slows dramatically. The first turning point (elbow) of the SSE curve suggests the "right" number of clusters.

This procedure was repeated for several combinations of spatial weights and number of clusters resulting in the global optimum solution.

3. Results and discussion

3.1. ETo clusters

The clustering of the parameters of the Parametric ETo model for the 3139 stations involved in the analysis, was implemented using the spectral clustering method of the Geoda software, which is free and open source software tool for spatial data analysis (Anselin, 2024).

SSE analysis along with elbow plots were performed for several combinations of spatial weights and number of clusters to acquire the global optimum clusters size. The weight for geometric distance ranged from 1 (100%) to 0.85 (85%), while the combined weight of the two parameters ranged from 0 (0%) to 0.15 (15%), respectively. These limits were realized after a preliminary investigation into the physical meaning and rationality of the produced clusters.

The combination that provided the smallest SSE and concurrently the smaller number of meaningful clusters, was 0.96 for the geometric distance and 0.02 for each one of the parameters. Thus, the objective function to be minimized (eq. (5)) was transformed to:

$$\min(\text{SSE}) = \sum_{i=1}^{k} \sum_{s}^{h} \Big[0.96 \, dist(s, c_i)^2 + 0.02 \, dist(a', a'_{ci})^2$$
$$+ 0.02 \, dist(c', c'_{ci})^2 \Big] \tag{7}$$

where s is the great circle distance between each station and the center of cluster i. Also, a' and c' are the parameters of the modified Parametric method (eq. (2) and c'_{ci}, a'_{ci} are the values of parameters at the centroid of cluster c_i ($1 \leq i \leq k$), that includes h stations.

The above detailed approach resulted in the identification of 45 unique ETo clusters (CL1 to CL45) across the globe, each representing
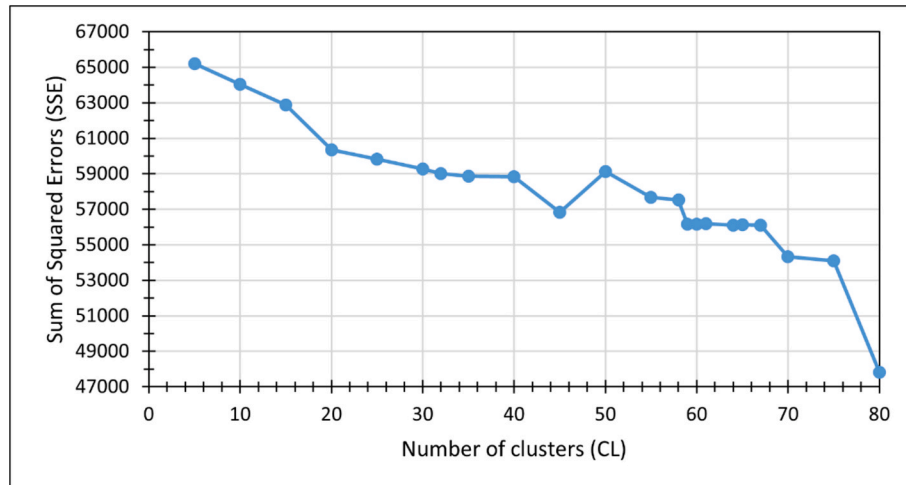

Fig. 1. Elbow plot depicting the optimum number of clusters (45) for the selected weights.
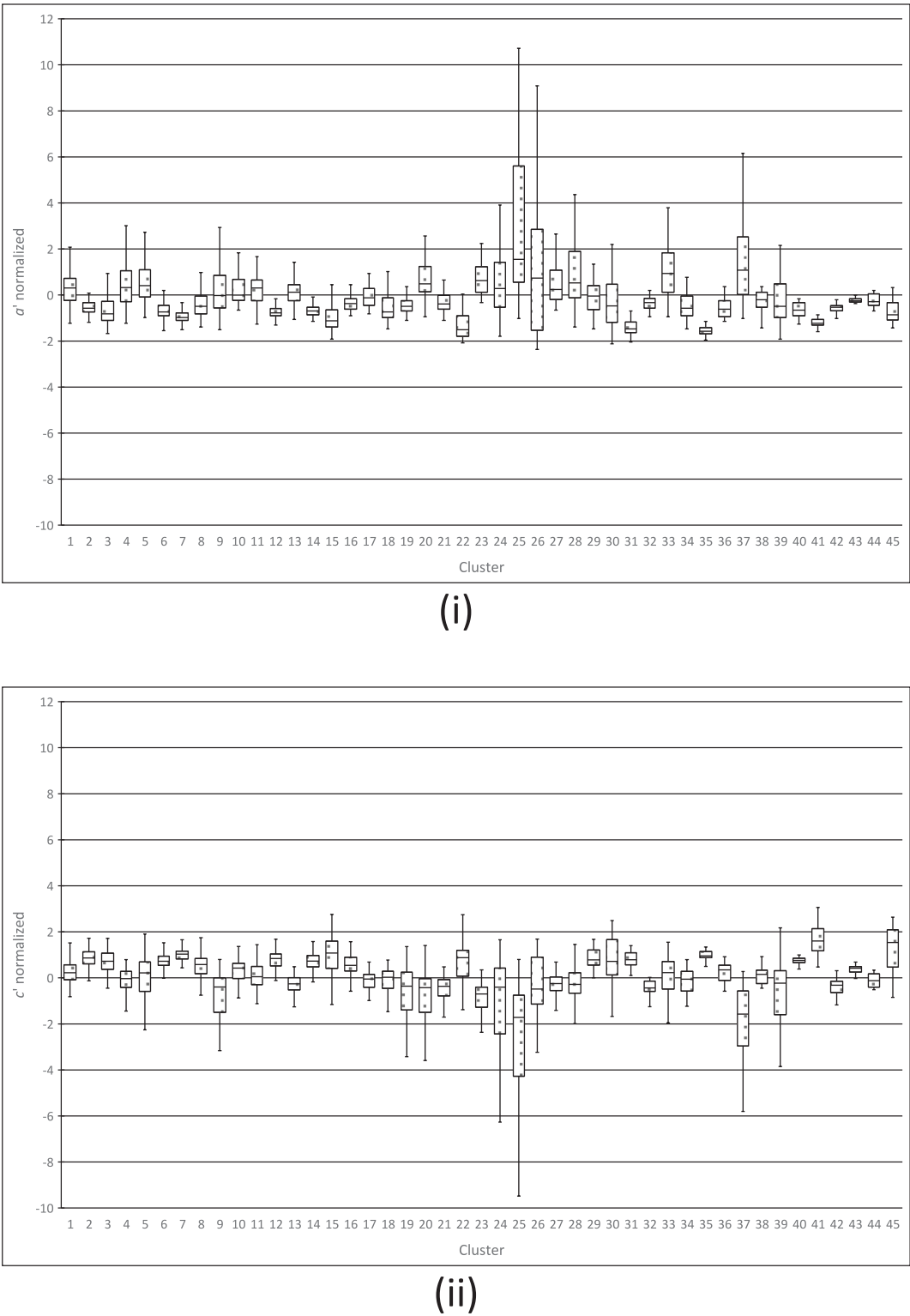
(i)

(ii)

Fig. 2. Boxplots of the normalized values of *a'* (i) and *c'* (ii) parameters within each cluster.

distinctive ETo characteristics. (Fig. 1). The clusters are numbered in descending order based on the number of contributing stations, with CL1 containing the most stations and CL45 the fewest.

To further assess the internal consistency of the identified clusters, we conducted a boxplot analysis of the normalized values of *a'* and *c'* within each cluster (Fig. 2). This allowed evaluation of the variability of the parameters while mitigating the effects of different numerical magnitudes. The results reveal that the clusters exhibit narrow

interquartile ranges (IQRs), indicating a high degree of internal consistency. A limited number of clusters, i.e. one or two, present wider IQRs for both parameters, suggesting some degree of intra-cluster variability. This intra-cluster variability is explained, given the evapotranspiration's complex nature that is influenced by multiple climatic and environmental factors in these regions, which are discussed further in Section 3.4.

Table 1 presents the values of the *a'* and *c'* parameters at each of the

Table 1
Values of the *a'* and *c'* parameters at the clusters centroids.

Cluster	*a'*	*c'*	Cluster	*a'*	*c'*	Cluster	*a'*	*c'*
1	7.384E-05	0.0210	16	5.950E-05	0.0229	31	3.380E-05	0.0249
2	5.401E-05	0.0263	17	6.618E-05	0.0170	32	5.998E-05	0.0147
3	5.054E-05	0.0255	18	5.523E-05	0.0169	33	1.032E-04	0.0188
4	8.164E-05	0.0176	19	5.624E-05	0.0125	34	5.689E-05	0.0172
5	8.127E-05	0.0177	20	8.493E-05	0.0116	35	2.945E-05	0.0270
6	5.005E-05	0.0254	21	5.989E-05	0.0143	36	6.039E-05	0.0181
7	4.382E-05	0.0278	22	4.209E-05	0.0234	37	1.267E-04	−0.0090
8	5.934E-05	0.0229	23	8.349E-05	0.0114	38	6.173E-05	0.0192
9	7.479E-05	0.0112	24	6.582E-05	0.0127	39	4.021E-05	0.0202
10	7.398E-05	0.0213	25	2.394E-04	−0.0456	40	5.091E-05	0.0246
11	7.184E-05	0.0197	26	5.900E-05	0.0253	41	3.741E-05	0.0323
12	4.885E-05	0.0256	27	8.171E-05	0.0160	42	5.200E-05	0.0159
13	7.005E-05	0.0162	28	9.492E-05	0.0168	43	6.269E-05	0.0222
14	5.121E-05	0.0249	29	6.547E-05	0.0268	44	6.114E-05	0.0177
15	4.515E-05	0.0265	30	6.170E-05	0.0252	45	4.971E-05	0.0253

45 clusters centers. Obviously, these values can be applied for the estimation of the ETo inside each cluster using eq. (2), providing means for direct application of the parametric method. The validation of the proposed values is presented in the corresponding section of this study.

Fig. 3 illustrates the spatial distribution of the 45 ETo clusters across the globe. The colored dots correspond to the meteorological stations involved in the clustering process, with their color indicating the specific cluster to which they belong.

Asia covering a large part of the globe, comprises from thirteen clusters: 1, 3, 10, 13, 17, 18, 21, 32, 38, 40, 41, 42 and 43. North America is dominated by clusters 11 and 29. Central and South America are covered by nine clusters: 8, 9, 22, 24, 25, 27, 30, 36, and 39; as Africa which is covered by clusters: 4, 15, 20, 23, 28, 31, 35, 37 and 44. Europe includes seven clusters: 2, 6, 7, 12, 14, 16 and 19, while Australia and Oceania include five clusters: 5, 26, 33, 34 and 45.

The clustering results proved that the spectral clustering approach effectively identified spatially coherent ETo clusters while incorporating the underlying evapotranspiration dynamics. Unlike traditional clustering methods, where the assigned weights directly influence the clustering structure, spectral clustering operates on a transformed representation of the data derived from the eigenvalues of the affinity matrix. This transformation mitigated the risk of any single variable dominating the clustering outcome, allowing for a more balanced integration of geographic and ETo attributes. The weighting scheme (0.96 for geometric distance, 0.02 each for *a'* and *c'*), determined through optimization, ensured that the resulting clusters maintained spatial coherence without neglecting the ETo parameters. Had the clustering been based solely on *a'* and *c'*, it would have resulted in geographically disconnected clusters, reducing the functionality for regional hydrological modeling and applications.

A detailed analysis of the distribution of the clusters is presented in the following sections.

3.2. Distribution of stations by clusters

Table 2 presents the distribution of stations across the 45 identified clusters. The number of stations per cluster ranges significantly, from as few as 14 stations (in clusters such as CL44 and CL45) to as many as 160 stations (in CL1). The distribution of stations across clusters reflects the global variability in both climatic zones and the availability of meteorological data. Larger clusters, in terms of the higher number of stations included, are associated with more widespread climatic conditions, such as arid or Mediterranean climates, which cover extensive geographic regions and have more available meteorological observations. Conversely, clusters with small number of stations included, indicate more localized or unique climatic zones, where distinct ETo dynamics prevail. These regions may require specialized water management strategies due to their specific ETo characteristics, as discussed in section

3.4 and depicted in Table 5.

Cluster 1 (160 stations) and cluster 2 (155 stations), are the largest clusters. These large clusters suggest broad climatic zones that are well-represented by a large number of weather stations. The size of these clusters indicates that the areas they represent have relatively homogeneous *a'* and *c'* values, i.e. similar ETo variability across the corresponding geographic regions. For instance, cluster 1, with stations from countries like Iran, Afghanistan, and Pakistan, likely represents arid and semi-arid climates where ETo is heavily driven by high temperatures and low precipitation. Cluster 2, dominated by stations in Türkiye and Greece, reflects regions with Mediterranean climates characterized by distinct dry and wet seasons. The significant number of stations in these clusters suggests that these regions have well-established meteorological data networks, allowing for robust cluster formation.

On the contrary, clusters 44 and 45, with only 14 stations each, represent much more geographically and climatically localized ETo conditions, such as certain temperate-tropical (Cluster 44) or oceanic regions (Cluster 45), which sets them apart from adjacent clusters.

Overall, the variation in cluster sizes is foreseeable, as regions with uniform climatic conditions naturally form larger clusters, while regions with more complex climatic influences create smaller, more distinct clusters. While station availability influences the number of stations per cluster, their distribution underscores the robustness of the clustering method, which captures both broad-scale climatic trends and localized ETo variations.

3.3. Distribution of clusters by country

Each of the 45 clusters identified through spectral clustering was analyzed in terms of geographic distribution, focusing on the countries contributing the most stations to each cluster. This analysis helps to understand how specific geographic regions align with the clusters based on ETo characteristics and climate factors, offering insights into regional variations in water needs and management practices.

Table 3 summarizes the top three countries contributing stations to each cluster, highlighting both the geographic and climatic diversity captured by the clustering method. The largest cluster, Cluster 1, is predominantly composed of stations from Iran (26.88%), Afghanistan (15.62%), and Pakistan (13.12%). This cluster represents a region characterized by a mixture of arid and semi-arid climates, which is consistent with the high reference evapotranspiration rates typically observed in these areas. The high percentage of stations from these countries suggests that this cluster may represent a climatic zone where ETo is strongly influenced by factors such as high temperatures and low humidity, particularly in regions with sparse vegetation and limited water availability.

Cluster 2 is primarily dominated by stations in Türkiye (86.45%), with smaller contributions from Greece (10.97%) and Iran (1.94%). The
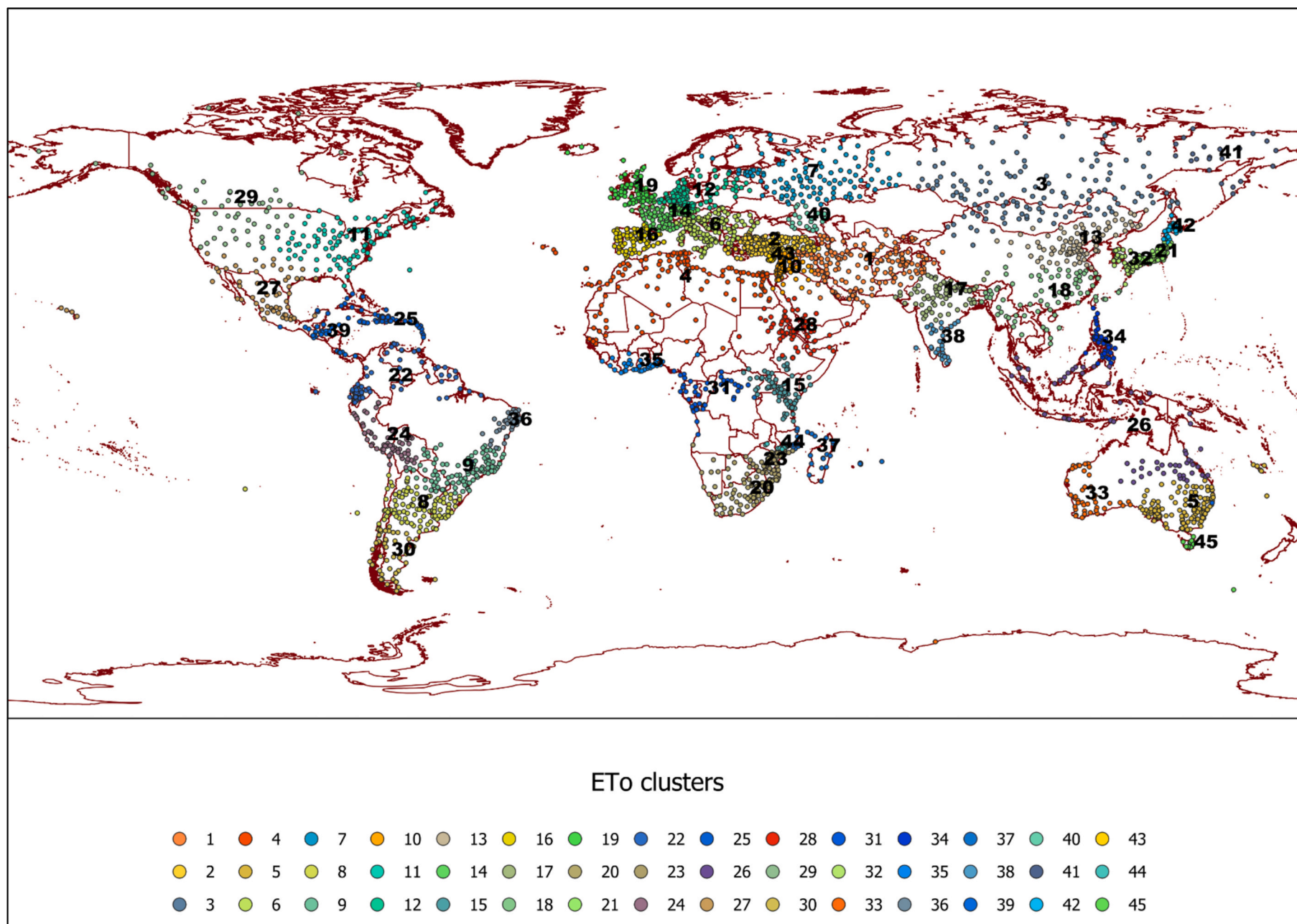
Fig. 3. Global ETo clustering of FAO ClimWat stations based on the parametric method.

Table 2
ETo Clusters and the corresponding number of stations included.

Cluster	Number of stations	Cluster	Number of stations	Cluster	Number of stations
1	160	16	80	31	46
2	155	17	77	32	43
3	136	18	75	33	43
4	123	19	74	34	37
5	120	20	73	35	33
6	118	21	71	36	33
7	114	22	70	37	33
8	111	23	69	38	30
9	105	24	67	39	29
10	99	25	66	40	22
11	96	26	65	41	22
12	95	27	63	42	22
13	88	28	61	43	16
14	86	29	58	44	14
15	81	30	46	45	14

distribution of this cluster reflects the Mediterranean climate, which is characterized by hot, dry summers and mild, wet winters. ETo in these regions is typically high during the summer months due to increased solar radiation and temperature, making this cluster significant for agricultural planning, particularly in the context of irrigation

management during drought periods.

The Russian Federation features prominently in several clusters, with the largest representation in Cluster 3, where 69.12% of the stations are in Russia, followed by Mongolia (17.65%). The dominance of Russian stations in this and other clusters (e.g., Cluster 7 with 86.84%) suggests that the methodology effectively captures the varied climatic zones within Russia, ranging from humid continental to subarctic climates. These regions are characterized by considerable seasonal variation in temperature, which directly impacts ETo rates. This variation in climatic conditions within a single country illustrates the robustness of the clustering approach, which is capable of distinguishing between distinct ETo patterns within large geographic regions.

In Africa, Cluster 15 is composed of stations from Kenya (39.51%), United Republic of Tanzania (19.75%), and Republic of Uganda (14.81%). This cluster likely represents tropical savanna and equatorial climates, where rainfall is seasonal, and temperatures remain consistently high throughout the year. The high ETo rates in these regions necessitate effective water resource management, particularly for agriculture, where irrigation is crucial during dry seasons.

Clusters 5, 33 and 45 are almost entirely composed of stations from Australia with percentages that range from 95.83% to 100%. Australia's presence in multiple clusters underscores the continent's wide range of climatic zones, from tropical and arid regions in the north and center to

Table 3
Clusters' station distribution across the three major contributing countries.

Cluster	Country 1	Country 2	Country 3
1	Iran (26.88%)	Afghanistan (15.62%)	Pakistan (13.12%)
2	Türkiye (86.45%)	Greece (10.97%)	Iran (1.94%)
3	Russian Federation (69.12%)	Mongolia (17.65%)	
4	Egypt (24.39%)	Algeria (17.89%)	Libya (13.82%)
5	Australia (95.83%)		
6	Italy (49.15%)	Romania (9.32%)	Croatia (7.63%)
7	Russian Federation (86.84%)	Estonia (8.77%)	Sweden (3.51%)
8	Argentina (61.26%)	Uruguay (15.32%)	Chile (13.51%)
9	Brazil (71.43%)	Argentina (11.43%)	Paraguay (11.43%)
10	Syria (34.34%)	Jordan (27.27%)	Lebanon (15.15%)
11	USA (79.17%)	Canada (19.79%)	
12	Germany (55.79%)	Netherlands (7.37%)	Poland (7.37%)
13	China (93.18%)	Russian Federation (5.68%)	
14	France (66.28%)	Germany (24.42%)	Italy (3.49%)
15	Kenya (39.51%)	Tanzania (19.75%)	Uganda (14.81%)
16	Spain (76.25%)	Portugal (17.50%)	France (5.00%)
17	India (68.83%)	Nepal (23.38%)	China (3.90%)
18	China (57.33%)	India (9.33%)	Vietnam (8.00%)
19	United Kingdom (77.03%)	Ireland (17.57%)	Iceland (2.70%)
20	South Africa (50.68%)	Mozambique (27.40%)	Lesotho (10.96%)
21	Japan (100.00%)		
22	Ecuador (47.14%)	Venezuela (20.00%)	Brazil (11.43%)
23	Mozambique (55.07%)	Zimbabwe (21.74%)	Botswana (10.14%)
24	Peru (59.70%)	Bolivia (37.31%)	
25	Haiti (27.27%)	Dominican Republic (16.67%)	Jamaica (10.61%)
26	Australia (50.77%)	Malaysia (20.00%)	Indonesia (18.46%)
27	Mexico (60.32%)		
28	Sudan (32.79%)	Saudi Arabia (18.03%)	Ethiopia (14.75%)
29	Canada (51.72%)		
30	Argentina (65.22%)	Chile (32.61%)	
31	Congo (50.00%)	Gabon (19.57%)	Congo (13.04%)
32	Japan (69.77%)		
33	Australia (100.00%)		
34	Philippines (97.30%)		
35	Cote D' Ivoire (33.33%)	Ghana (33.33%)	Benin (9.09%)
36	Brazil (100.00%)		
37	Madagascar (39.39%)	Mozambique (33.33%)	Mauritius (15.15%)
38	India (76.67%)	Sri Lanka (23.33%)	
39	Guatemala (34.48%)	Honduras (34.48%)	Costa Rica (13.79%)
40	Russian Federation (100.00%)		
41	Russian Federation (100.00%)		
42	Japan (95.45%)		
43	Cyprus (87.50%)	Turkey (12.50%)	
44	Mozambique (78.57%)	Malawi (21.43%)	
45	Australia (100.00%)		

temperate zones in the south. The clustering results reflect the diversity of ETo patterns in Australia, with high evapotranspiration rates in arid and semi-arid regions driving water management practices focused on conservation and efficient irrigation.

Interestingly, some clusters reflect more isolated station distributions. For instance, Cluster 27 is dominated by stations from Mexico (60.32%), emphasizing the importance of local climatic factors in determining ETo patterns. This cluster may represent the distinct climatic conditions found in Mexico's arid and semi-arid regions, where evapotranspiration is a critical factor in agricultural water use. Another example is Cluster 34 which contains stations from Philippines (97.30%) depicting the unique and diverse tropical maritime climate of the archipelago.

Overall, the geographic distribution of stations within each cluster aligns well with the expected climatic patterns of the regions, providing strong validation of the clustering approach. The presence of distinct country groupings within each cluster indicates that the method successfully captures both local and regional variability in ETo, contributing to a more nuanced understanding of global evapotranspiration dynamics.

3.4. Analogy to Köppen-Geiger climate classification

The Köppen-Geiger climate classification (Beck et al., 2018; Kottek et al., 2006) (Table 4), a widely recognized system for delineating global climate zones based on temperature, precipitation, and their seasonal patterns, was used to correlate these ETo clusters with specific climate types. To assess the correlation between the clusters and specific climate zones, an analysis was conducted to determine the dominant Köppen-Geiger climate types within each cluster. This approach allows the assessment of the relationship between the proposed ETo clusters and climatic conditions across the globe.

Fig. 4 illustrates the distribution of stations across Köppen-Geiger classes, demonstrating a strong alignment between large number of stations and specific climate zones. On the other hand, the small number of stations noticed at specific Köppen-Geiger classes can be justified from the small number or even absence of monitoring networks in those regions.

Fig. 5 illustrates the ETo clusters, as points at the stations locations, compared to Köppen-Geiger classes across the globe. It is apparent that the proposed clusters respect the distinguished climate patterns, since in most of the cases they are allocated within specific climate zones.

Table 4
Köppen-Geiger climate classification.

| Class	Code	Classification	Class	Code	Classification
1	Af	Tropical, rainforest	17	Dsa	Cold, dry summer, hot summer
2	Am	Tropical, monsoon	18	Dsb	Cold, dry summer, warm summer
3	Aw	Tropical, savannah	19	Dsc	Cold, dry summer, cold summer
4	BWh	Arid, desert, hot	20	Dsd	Cold, dry summer, very cold winter
5	BWk	Arid, desert, cold	21	Dwa	Cold, dry winter, hot summer
6	BSh	Arid, steppe, hot	22	Dwb	Cold, dry winter, warm summer
7	BSk	Arid, steppe, cold	23	Dwc	Cold, dry winter, cold summer
8	Csa	Temperate, dry summer, hot summer	24	Dwd	Cold, dry winter, very cold winter
9	Csb	Temperate, dry summer, warm summer	25	Dfa	Cold, no dry season, hot summer
10	Csc	Temperate, dry summer, cold summer	26	Dfb	Cold, no dry season, warm summer
11	Cwa	Temperate, dry winter, hot summer	27	Dfc	Cold, no dry season, cold summer
12	Cwb	Temperate, dry winter, warm summer	28	Dfd	Cold, no dry season, very cold winter
13	Cwc	Temperate, dry winter, cold summer	29	ET	Polar, tundra
14	Cfa	Temperate, no dry season, hot summer	30	EF	Polar, frost
15	Cfb	Temperate, no dry season, warm summer			
16	Cfc	Temperate, no dry season, cold summer			
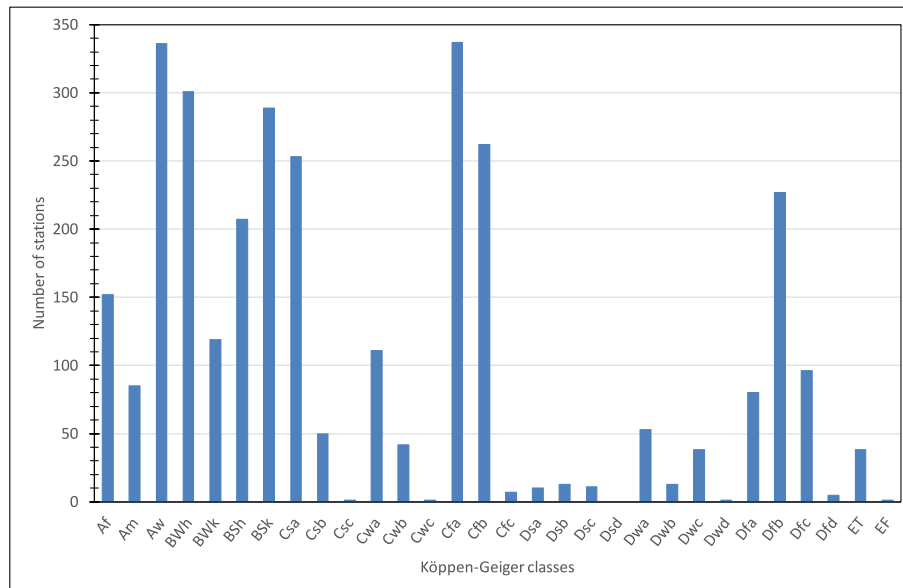
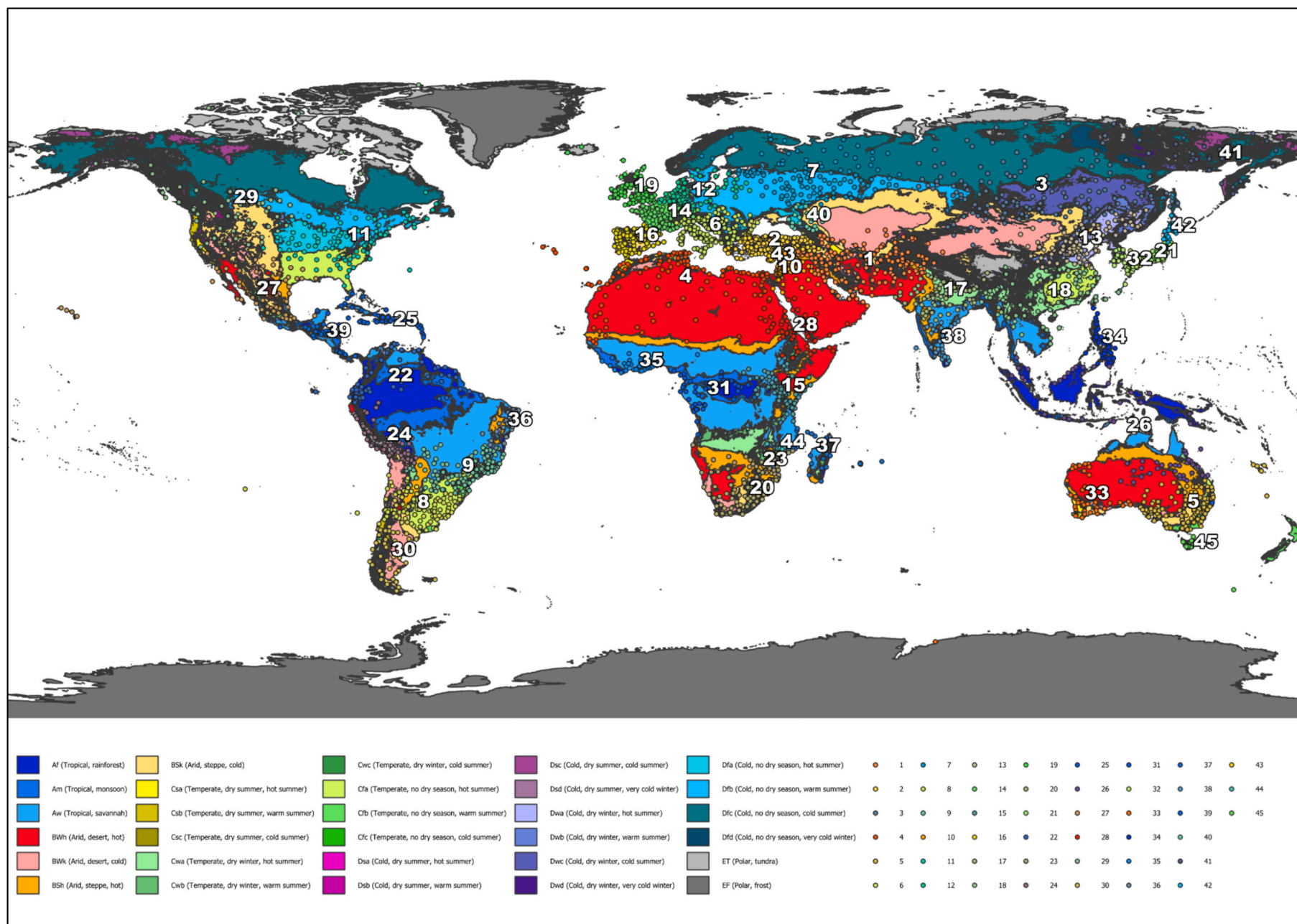**Fig. 4.** Distribution of 3139 stations among the Köppen-Geiger climate classification.

Fig. 5. Global ETo clustering compared to Köppen-Geiger classes.

Table 5

Distribution of the Köppen-Geiger climate classes among the 45 clusters.

Cluster	Köppen-Geiger Class										
25	Af (36.4%)	Aw (36.4%)	Am (13.6%)	BSh (6.1%)	Cfb (4.5%)	Cwb (1.5%)	Cfa (1.5%)				
26	Af (38.5%)	BSh (23.1%)	BWh (13.8%)	Aw (12.3%)	Am (7.7%)	Cwa (1.5%)	Cwb (1.5%)	Cfa (1.5%)			
34	Af (56.8%)	Am (24.3%)	Aw (18.9%)								
22	Aw (30.0%)	Af (27.1%)	Am (18.6%)	BSh (10.0%)	Cfb (7.1%)	BWh (2.9%)	ET (2.9%)	BSk (1.4%)			
9	Aw (36.2%)	Cfa (30.5%)	Cwa (13.3%)	Cfb (6.7%)	Am (3.8%)	Af (2.9%)	BSh (2.9%)	BSk (1.9%)	Cwb (1.9%)		
36	Aw (39.4%)	BSh (36.4%)	Am (15.2%)	Af (9.1%)							
37	Aw (42.4%)	Af (27.3%)	Am (9.1%)	Cwb (9.1%)	BSh (6.1%)	Csb (3%)	Cwa (3%)				
31	Aw (43.5%)	Af (32.6%)	Am (19.6%)	BWh (2.2%)	BSh (2.2%)						
39	Aw (44.8%)	Af (27.6%)	Am (17.2%)	Cfb (6.9%)	Cwb (3.4%)						
15	Aw (46.9%)	Cfb (14.8%)	BSh (12.3%)	Af (8.6%)	Csb (7.4%)	Cwb (6.2%)	Am (3.7%)				
35	Aw (75.8%)	Am (21.2%)	Af (3%)								
38	Aw (76.7%)	BSh (16.7%)	Af (3.3%)	Cfb (3.3%)							
23	BSh (37.7%)	Aw (33.3%)	BWh (10.1%)	Cwa (10.1%)	Cwb (5.8%)	BSk (2.9%)					
29	BSk (31.0%)	Dfb (20.7%)	Dfc (10.3%)	ET (8.6%)	Csb (6.9%)	Cfb (6.9%)	BWk (5.2%)	Dfa (5.2%)	Csa (1.7%)	Dsc (1.7%)	Dwa (1.7%)
16	BSk (33.8%)	Csa (27.5%)	Csb (16.3%)	Cfb (15.0%)	BSh (3.8%)	BWh (2.5%)	Cfa (1.3%)				
27	BWh (19.0%)	BSh (19.0%)	Aw (15.9%)	BSk (15.9%)	Cfa (12.7%)	BWk (4.8%)	Cwb (4.8%)	Cwa (3.2%)	Af (1.6%)	Am (1.6%)	Cfb (1.6%)
24	BWh (23.9%)	ET (20.9%)	Af (14.9%)	Cwb (11.9%)	Aw (9.0%)	BSk (7.5%)	Am (4.5%)	BWk (3.0%)	BSh (1.5%)	Cwa (1.5%)	Cwc (1.5%)
1	BWh (30.0%)	BWk (21.3%)	BSk (18.1%)	Csa (12.5%)	BSh (10.6%)	Dsa (2.5%)	Cwa (1.3%)	Cfa (1.3%)	ET (1.3%)	Dsb (0.6%)	Dfb (0.6%)
33	BWh (44.2%)	BWk (11.6%)	BSh (11.6%)	BSk (11.6%)	Csa (9.3%)	Csb (9.3%)	EF (2.3%)				
4	BWh (63.4%)	Csa (17.1%)	BSh (11.4%)	BSk (4.1%)	Aw (3.3%)	BWk (0.8%)					
28	BWh (73.8%)	BSh (13.1%)	Aw (6.6%)	BSk (4.9%)	Csb (1.6%)						
30	BWk (30.4%)	BSk (23.9%)	Csb (19.6%)	Cfb (10.9%)	Cfc (8.7%)	ET (4.3%)	Csc (2.2%)				
5	Cfa (25.0%)	BSk (22.5%)	BWh (14.2%)	Cfb (12.5%)	BSh (10.8%)	BWk (5.0%)	Csb (3.3%)	Csa (2.5%)	Am (1.7%)	Aw (1.7%)	Af (0.8%)
18	Cfa (41.3%)	Cwa (24.0%)	Aw (17.3%)	Am (5.3%)	Af (4.0%)	BSh (4.0%)	Dwc (2.7%)	Dwb (1.3%)			
8	Cfa (50.5%)	BWk (11.7%)	Cwa (9.0%)	BSk (7.2%)	BSh (5.4%)	Cfb (4.5%)	Csa (3.6%)	Csb (3.6%)	BWh (1.8%)	ET (1.8%)	Am (0.9%)
32	Cfa (79.1%)	Dwa (14.0%)	Dfa (4.7%)	Cwa (2.3%)							
21	Cfa (81.7%)	Dfa (16.9%)	Dfb (1.4%)								
12	Cfb (50.5%)	Dfc (1.1%)									
14	Cfb (64.0%)	Dfb (17.4%)	Csa (11.6%)	BSk (2.3%)	Cfa (2.3%)	Dfc (1.2%)	ET (1.2%)				
45	Cfb (85.7%)	BSk (7.1%)	ET (7.1%)								
19	Cfb (93.2%)	Cfc (4.1%)	Dsc (1.4%)	Dfc (1.4%)							
10	Csa (45.5%)	BWh (26.3%)	BSk (14.1%)	BSh (12.1%)	BWk (1.0%)	Dsb (1.0%)					
2	Csa (45.8%)	BSk (27.1%)	Cfa (8.4%)	Dsb (6.5%)	Dsa (3.9%)	Dfb (3.2%)	Cfb (2.6%)	BWk (1.3%)	Csb (1.3%)		
43	Csa (62.5%)	BSh (31.3%)	Csb (6.3%)								
17	Cwa (48.1%)	Aw (18.2%)	BSh (16.9%)	BWh (11.7%)	ET (2.6%)	BSk (1.3%)	Cwb (1.3%)				
44	Cwa (57.1%)	Aw (28.6%)	Am (14.3%)								
20	Cwb (17.8%)	Aw (16.4%)	BSh (13.7%)	BWh (11.0%)	BWk (11.0%)	BSk (11.0%)	Cwa (8.2%)	Cfa (5.5%)	Cfb (2.7%)	Csb (1.4%)	Dwc (1.4%)
6	Dfa (1.7%)	ET (1.7%)	BSk (13.6%)	Dfb (21.2%)	Cfa (26.3%)	Csa (35.6%)					
11	Dfa (39.6%)	Cfa (31.3%)	Dfb (24.0%)	Dfc (3.1%)	Af (1.0%)	BSk (1.0 %)					
40	Dfa (40.9%)	BSk (27.3%)	BWk (13.6%)	Cfa (9.1%)	Dfb (9.1%)						
42	Dfb (54.5%)	Dfa (45.5%)									
7	Dfb (61.4%)	Dfc (31.6%)	BSk (3.5%)	Dfa (3.5%)							
3	Dfc (29.4%)	Dwc (25.0%)	BWk (11.8%)	BSk (11.0%)	Dfb (10.3%)	Dwb (3.7%)	Dwa (2.9%)	Dfd (2.2%)	Dsc (1.5%)	Dsb (0.7%)	Dwd (0.7%)
41	Dfc (36.4%)	Dsc (31.8%)	ET (18.2%)	Dfd (9.1%)	Dwc (4.5%)						
13	Dwa (47.7%)	BSk (29.5%)	BWk (9.1%)	Dwb (8.0%)	Cwa (3.4%)	Cfa (1.1%)	Dfb (1.1%)				

Table 5 provides a detailed breakdown of the Köppen-Geiger classes within each ETo cluster. It is obvious that each cluster is primarily composed of a few dominant climate types, with additional classifications contributing only minor fractions. This suggests that while ETo clusters and climate zones are not identical, they share meaningful climatic relationships, reinforcing the validity of the clustering approach. This presence of multiple Köppen-Geiger climate zones within an ETo cluster reflects the continuity of climate transitions and localized geographic effects on ETo dynamics.

For example, Cluster 1 is primarily dominated by hot desert climates (BWh) and cold semi-arid climates (BSk), with some Mediterranean influences (Csa), reflecting the arid, dry conditions typical of desert regions with occasional Mediterranean climates. In Cluster 2, Mediterranean climate (Csa) is the dominant type, followed by cold semi-arid (BSk) and humid subtropical (Cfa) climates, indicating dry summers, wet winters, and moderate temperature variation. Cluster 3 is characterized by cold climates, particularly subarctic (Dfc) and monsoon-influenced climates (Dwc), with the presence of desert climates (BWk) highlighting the cold and arid regions.

Cluster 4 is dominated by hot desert climates (BWh), indicating extremely arid conditions with minimal rainfall and minor Mediterranean influences (Csa). In Cluster 5, humid subtropical climates (Cfa) dominate, followed by cold semi-arid (BSk) and hot desert (BWh) climates, representing warmer, wetter climates alongside dry desert conditions. Cluster 6 is primarily Mediterranean (Csa), with humid subtropical (Cfa) and cold semi-arid (BSk) influences, reflecting warm, dry summers and wet winters. In Cluster 7, cold, subarctic climates (Dfb, Dfc) dominate, representing regions with long, harsh winters and short summers. Cluster 8 is dominated by humid subtropical climates (Cfa), followed by desert (BWk) and semi-arid (BSk) climates, indicating a mix of warm, humid conditions alongside arid regions. Cluster 9 features tropical savanna (Aw) and humid subtropical (Cfa) climates, with a diverse mix of monsoon-influenced (Cwa) and semi-arid (BSh) climates.

Cluster 10 is dominated by Mediterranean climates (Csa) and hot desert (BWh), reflecting hot, dry summers and arid desert conditions. Cluster 11 is characterized by cold continental (Dfa) and humid subtropical (Cfa) climates, reflecting colder, wetter conditions with seasonal variation. In Cluster 12, oceanic (Cfb) and cold continental (Dfb) climates dominate, indicating regions with moderate temperatures and significant precipitation throughout the year. Cluster 13 is primarily cold monsoon (Dwa) and cold semi-arid (BSk), reflecting significant temperature variation and semi-arid conditions. Cluster 14 is dominated by oceanic (Cfb) climates, with Mediterranean (Csa) and cold semi-arid (BSk) influences, indicating wet, mild conditions with dry summers. Cluster 15 features tropical savanna (Aw) as the primary climate, followed by oceanic (Cfb) and semi-arid (BSh) climates, reflecting tropical wet-dry conditions and temperate influences. In Cluster 16, cold semi-arid (BSk) and Mediterranean (Csa) climates dominate, reflecting dry, arid conditions with seasonal temperature variation. Cluster 17 is characterized by humid subtropical (Cwa) and tropical savanna (Aw) climates, representing regions with significant seasonal rainfall and warm temperatures. Cluster 18 is dominated by humid subtropical climates (Cfa), with tropical savanna (Aw) and monsoon influences (Am), indicating warm, wet conditions.

Cluster 19 is primarily oceanic (Cfb), representing regions with mild, wet conditions throughout the year. Cluster 20 is dominated by subtropical highland (Cwb) and tropical savanna (Aw) climates, reflecting mild temperatures and significant rainfall. Cluster 21 is primarily humid subtropical (Cfa), representing warm, wet conditions typical of subtropical regions. In Cluster 22, tropical savanna (Aw) and tropical rainforest (Af) dominate, indicating regions with consistent warmth and significant rainfall. Cluster 23 is distinguished by semi-arid (BSh) and tropical savanna (Aw) climates, representing hot, dry conditions with some seasonal rainfall. Cluster 24 is dominated by hot desert (BWh) and tundra (ET), reflecting extremely arid conditions with occasional colder, high-altitude regions.

Cluster 25 is characterized by tropical rainforest (Af) and tropical savanna (Aw), reflecting warm, wet conditions typical of equatorial regions. Cluster 26 is dominated by tropical rainforest climates (Af), followed by semi-arid (BSh) and desert (BWh) climates, indicating a mix of warm, wet and arid conditions. The presence of humid tropical areas alongside drier semi-arid regions reflects the heterogeneity in ETo patterns, encapsulated through the spectral clustering of the two parameters. Cluster 27 is dominated by hot desert (BWh) and semi-arid (BSh) climates, reflecting arid desert regions with limited rainfall.

In Cluster 28, hot desert climates (BWh) dominate, reflecting extreme arid conditions with little to no rainfall. Cluster 29 features cold semi-arid (BSk) and cold continental (Dfb, Dfc) climates, reflecting dry, cold conditions with harsh winters. Cluster 30 is dominated by cold desert (BWk) and cold semi-arid (BSk), reflecting arid regions with cold, dry conditions.

Cluster 31 is primarily tropical savanna (Aw) and tropical rainforest (Af), representing warm, wet conditions with seasonal rainfall. Cluster 32 is dominated by humid subtropical (Cfa) climates, reflecting warm, wet conditions throughout the year. Cluster 33 is distinguished by hot desert (BWh) and cold desert (BWk) climates, reflecting arid conditions with minimal rainfall. In Cluster 34, tropical rainforest climates (Af) dominate, indicating warm, wet conditions typical of equatorial regions. Cluster 35 is primarily tropical savanna (Aw), reflecting regions with seasonal rainfall and warm temperatures. Cluster 36 features tropical savanna (Aw) and semi-arid (BSh) climates, representing warm, dry conditions with some seasonal rainfall. Cluster 37 is dominated by tropical savanna (Aw), with tropical rainforest (Af) and subtropical highland (Cwb) influences, indicating a mix of wet, tropical and temperate conditions. Cluster 38 is primarily tropical savanna (Aw), followed by semi-arid (BSh), indicating warm, dry conditions. Cluster 39 features tropical savanna (Aw) and tropical rainforest (Af) climates, with some temperate influences (Cfb, Cwb), reflecting a mix of wet and dry tropical conditions.

Cluster 40 is characterized by cold continental (Dfa) and semi-arid (BSk) climates, indicating cold, dry conditions with significant seasonal variation. Cluster 41 is dominated by subarctic (Dfc) and tundra (ET) climates, reflecting regions with long, harsh winters and minimal rainfall. Cluster 42 is primarily cold continental climates (Dfb, Dfa), reflecting cold, wet conditions with significant seasonal variation. In Cluster 43, Mediterranean (Csa) and semi-arid (BSh) climates dominate, indicating dry summers and moderate temperatures. Cluster 44 is primarily humid subtropical (Cwa) and tropical savanna (Aw), reflecting warm, wet conditions with seasonal rainfall. Finally, Cluster 45 is dominated by oceanic (Cfb) climates, indicating mild wet conditions throughout the year.

3.5. Clustered ETo validation

Even though this study depicts the global clustering of monthly ETo, thus the areas for calibrating methods of ETo estimation, it also provides a way to compute ETo using the clustered parameters across the globe in a different, simpler way than the one presented in Tegos et al., 2017. This can be achieved using the parameters' values that correspond to the cluster that includes each area of interest, rather than performing calculations in GIS environment. Obviously, there should be a tradeoff between accuracy and ease of use, however the performance is acceptable for hydrological and engineering tasks in monthly scale as depicted in Table 6.

The validation dataset comprised from 30 stations from countries with different hydroclimatic regimes (USA, China, Spain, Germany, Ireland, Greece and Australia), for which we obtained full time series of the required meteorological data, at the monthly scale, from various data sources (Du et al., 2016; Gentilucci et al., 2020; Mamassis et al., 2021; Tegos et al., 2017, 2015b; Webb, 2010).

For this purpose, we implemented several accuracy measures such as the Nash-Sutcliffe Efficiency (NSE), the mean absolute error (MAE), the

Table 6
Statistical indexes for the monthly ETo validation dataset.

#	Station	Country	Validation period	NSE	MAE (mm)	NMAE (%)	r^2	NRMSE (%)
1	Camerino	Italy	January 2010 − December 2019	0.940	8.6	12.2	0.975	6.9
2	Ancona	Italy	January 2010 − December 2019	0.961	7.7	9.6	0.977	5.8
3	Macerata	Italy	January 2010 − December 2019	0.927	11.2	13.7	0.969	7.4
4	HRB M2	China	January 2003 − December 2014	0.750	22.1	23.5	0.958	16.4
5	Ft. Lauderdale	USA	February 2004 − December 2018	0.686	13.2	12.2	0.806	15.8
6	Eastern Adams County	USA	February 2003 − November 2016	0.556	18.4	23.3	0.878	15.4
7	Buntigville	USA	January 2003 − December 2016	0.764	24.1	25.7	0.942	12.4
8	Davis	USA	January 2003 − December 2016	0.919	14.2	12.6	0.945	8.2
9	McArthur	USA	January 2003 − December 2016	0.822	18.5	21.0	0.968	12.2
10	Temecula	USA	January 2003 − December 2016	0.839	12.0	11.2	0.890	12.2
11	Manteca	USA	January 2003 − December 2016	0.943	11.8	10.3	0.948	6.8
12	Tulelake FS	USA	January 2003 − December 2016	0.773	18.8	23.1	0.963	13.7
13	Aachen-Hörn	Germany	January 2003 − May 2011	0.924	7.4	12.8	0.952	7.2
14	Angermunde	Germany	January 2003 − May 2013	0.656	18.3	33.2	0.681	15.1
15	Bremen	Germany	January 2003 − June 2013	0.944	6.2	11.6	0.948	6.0
16	Dresden	Germany	January 2003 − June 2013	0.884	10.2	16.5	0.930	8.3
17	Dusseldorf	Germany	January 2003 − June 2013	0.922	8.0	13.0	0.953	6.7
18	Frankfurt	Germany	January 2003 − June 2013	0.916	9.0	13.7	0.967	8.0
19	Hamburg	Germany	January 2003 − June 2013	0.945	6.1	11.4	0.950	6.0
20	Karlshue	Germany	January 2003 − September 2008	0.863	12.2	17.7	0.949	10.5
21	München	Germany	January 2003 − May 2013	0.715	15.7	27.5	0.720	14.9
22	Alicante	Spain	January 2003 − October 2009	0.887	12.3	11.3	0.964	10.4
23	Zaragoza	Spain	January 2003 − November 2009	0.705	29.7	25.9	0.960	16.2
24	Mace Head	Ireland	January 2011 − January 2022	0.930	5.2	11.0	0.946	7.3
25	Kostakioi	Greece	March 2015 − September 2019	0.766	13.1	15.7	0.915	13.9
26	Tunnack Fire Station	Australia	January 2003 − December 2016	0.663	20.3	25.3	0.954	15.2
27	Adelaide Airport	Australia	January 2003 − December 2016	0.772	24.3	17.9	0.973	14.3
28	Sydney Airport AMO	Australia	January 2003 − December 2016	0.655	19.5	14.6	0.831	15.3
29	Alice Springs Airport	Australia	January 2009 − December 2014	0.878	19.9	10.6	0.903	10.2
30	Newman_Aero	Australia	January 2009 − December 2016	0.788	24.1	12.1	0.890	12.3

normalized mean absolute error (NMAE), the coefficient of determination (r^2) and the normalized root mean square error (NRMSE) (Koutsoyiannis, 2025; Tegos et al., 2017; Willmott, 1982, 1981):

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(\overline{o} - o_i)^2} \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - o_i| \tag{9}$$

$$\text{NMAE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|p_i - o_i|}{\overline{o}} \tag{10}$$

$$r^2 = \left(\frac{\sum_{i=1}^{n}(o_i - \overline{o})(p_i - \overline{p})}{\sqrt{\sum_{i=1}^{n}(o_i - \overline{o})^2}\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}}\right)^2 \tag{11}$$

$$\text{NRMSE} = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2\right]^{1/2}}{\overline{o}} \tag{12}$$

where n is the number of predictions / observations, o_i is the ith observation, p_i is the ith prediction, $\overline{o}$ is the observations average while $\overline{p}$ is the predictions average.

Most stations demonstrate excellent performance with high NSE values, low MAE and NMAE, and high r^2, indicating that the variance of the observed values has been accounted along with small bias, thus there is agreement between them and the estimated ETo values (Koutsoyiannis, 2025). For instance, Camerino and Ancona in Italy achieve NSE values of 0.940 and 0.961, respectively, with low MAE (8.6 mm and 7.7 mm), r^2 values above 0.97. Similarly, stations like Manteca (USA), Hamburg (Germany), and Mace Head (Ireland) show NSE values above 0.930, with relatively low errors, confirming the method's

robustness in temperate climates.

Only a few stations exhibit lower performance, i.e. Zaragoza (Spain) shows a NSE of 0.705 and a MAE of 29.7 mm, although the r^2 remains high at 0.960, indicating some consistency since the variance of the observed values was acknowledged but with large bias. Eastern Adams County (USA) and Angermunde (Germany) also have low NSE values, i. e. 0.556 and 0.656 respectively, with high MAE and NMAE values, suggesting more significant deviations between estimated and observed ETo.

Overall, the validation procedure indicated that the use of clustered values of the two parameters provides a balance between ease of use and acceptable accuracy for practical applications across diverse climates.

4. Conclusions

This study introduces a novel approach to global reference evapotranspiration (ETo) estimation using spectral clustering techniques, which groups regions based on shared ETo dynamics derived from the parametric model variables (a' and c'). Through this approach, 45 distinct clusters were identified globally, with each cluster representing unique climatic and evapotranspiration characteristics.

The correspondence between the ETo clusters and Köppen-Geiger classification makes underscores the coherence of the clustering approach, as it groups areas with similar climatic behaviors into clearly defined clusters, similar to how the Köppen-Geiger system categorizes climates based on temperature and precipitation. This analogy provides a structured framework that connects ETo dynamics with global climate patterns.

A critical benefit of this clustering approach is its application to calibrating other ETo estimation methods. The spatial extent of the clusters can be used to guide and calibrate other models, such as temperature- and radiation-based models. The delineation of these clusters provides a geographical framework for grouping regions that exhibit similar ETo behaviors, allowing other models to be adjusted or fine-tuned based on the spatial extent of these coherent clusters. This

significantly simplifies the calibration process, as users can focus on predefined areas that have similar evapotranspiration characteristics without needing to account for the intricate variations across large geographic regions.

While the pre-calculated parameters derived from each cluster obviously provide means for estimating ETo using the Parametric method, the spatial boundaries of the clusters serve as a tool for improving the accuracy and efficiency of other methods. By using the spatial extent of these clusters, researchers can avoid recalibrating models for individual locations, instead applying a broader, more efficient approach that accounts for regional climatic variability. This reduces the computational burden and technical complexity of ETo estimation, especially in areas with limited meteorological data or resources.

The use of the pre-calculated parameters for each cluster obviously provides means for estimating ETo using the Parametric method. The validation results underscore the robustness of the clustered Parametric approach across a wide range of climatic conditions. The analysis of 30 stations from diverse hydroclimatic regions, such as the USA, China, Spain, Germany, Ireland, Greece, and Australia, demonstrated strong performance with high NSE values and high correlations (r^2) between observed and estimated ETo values. In many cases the achieved NSE values were above 0.9, indicating the method's suitability for hydrological and engineering tasks on a monthly scale. The validation also revealed limitations in regions with complex microclimates, such as Eastern Adams County in USA, where the clustered method showed a decrease in performance. This suggests that while the clusters are highly effective for larger regions, additional refinements may be required for areas with unique local conditions or laying close to the clusters limits.

The trade-off between accuracy and ease of use is one of the most significant advantages of the clustered approach. By providing both a simplified method for ETo estimation and a framework for calibrating other models, this study offers a solution that balances precision with practicality. The clustering approach reduces the need for detailed, location-specific data, making it an ideal tool for large-scale water resource management, agricultural planning, and climate resilience strategies, especially in regions where high-resolution datasets are scarce or unavailable.

The dataset of 3139 stations, including geographical information, parameter values for *a'* and *c'*, model performance metrics, and clustering details, is available for download from https://www.itia.ntua.gr/en/docinfo/2523/. This dataset aims to support further research in testing and refining the clustering results, particularly in regions with complex microclimates, while also promoting the applicability of the findings from this study.

CRediT authorship contribution statement

Nikolaos Malamos: Writing – original draft, Supervision, Methodology, Conceptualization. **Aristotelis Tegos:** Writing – review & editing, Validation, Investigation. **George Bourantas:** Visualization, Software, Formal analysis, Data curation. **Christos Chalvantzis:** Writing – original draft. **Demetris Koutsoyiannis:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have uploaded the complete dataset to our data repository and included a direct link for download, in the last paragraph of the Conclusions section

References

Akhavan, S., Mousabeygi, F., Peel, M.C., 2018. Assessment of eight reference evapotranspiration (ET$_0$) methods considering Köppen climate class in Iran. Hydrol. Sci. J. 63, 1468–1481. https://doi.org/10.1080/02626667.2018.1513654.

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. FAO, Rome.

Almorox, J., Quej, V.H., Martí, P., 2015. Global performance ranking of temperature-based approaches for evapotranspiration estimation considering Köppen climate classes. J. Hydrol. 528, 514–522. https://doi.org/10.1016/j.jhydrol.2015.06.057.

Anselin, L., 2024. An Introduction to Spatial Data Science with GeoDa: Volume 1: Exploring Spatial Data, 1st ed. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781003274919.

Antonopoulos, V.Z., Antonopoulos, A.V., 2017. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. Comput. Electron. Agric. 132, 86–96. https://doi.org/10.1016/j.compag.2016.11.011.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future köppen-geiger climate classification maps at 1-km resolution. Sci. Data 5, 1–12. https://doi.org/10.1038/sdata.2018.214.

Bellido-Jiménez, J.A., Estévez, J., García-Marín, A.P., 2022. A regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain. Agric Water Manag 274, 107955. https://doi.org/10.1016/j.agwat.2022.107955.

Demertzi, K., Papamichail, D., Aschonitis, V., Miliaresis, G., 2016. Hydroclimatic analysis of Greece using multi-parametric clustering of monthly precipitation and reference crop evapotranspiration. European Water 141–155.

Di Nunno, F., Granata, F., 2023. Future trends of reference evapotranspiration in Sicily based on CORDEX data and Machine Learning algorithms. Agric Water Manag 280, 108232. https://doi.org/10.1016/j.agwat.2023.108232.

Doorenbos, J., Pruitt, W.O., 1977. Guidelines for predicting crop water requirements, FAO Irrigation and Drainage Paper.

Du, C., Yu, J., Wang, P., Zhang, Y., 2016. Reference evapotranspiration changes: sensitivities to and contributions of meteorological factors in the Heihe River Basin of Northwestern China (1961–2014). Adv. Meteorol. 2016, 1–17. https://doi.org/10.1155/2016/4143580.

Earle, M.A., 2005. Vector solutions for great circle navigation. J. Navigation 58, 451–457. https://doi.org/10.1017/S0373463305003358.

Ferreira, L.B., Da Cunha, F.F., Fernandes Filho, E.I., 2022. Exploring machine learning and multi-task learning to estimate meteorological data and reference evapotranspiration across Brazil. Agric Water Manag 259, 107281. https://doi.org/10.1016/j.agwat.2021.107281.

Food and Agriculture Organization of the United Nations (FAO), 1993. CLIMWAT for CROPWAT: A climatic database for irrigation planning and management.

Gentilucci, M., Materazzi, M., Pambianchi, G., Burt, P., Guerriero, G., 2020. Temperature variations in Central Italy (Marche region) and effects on wine grape production. Theor. Appl. Climatol. 140, 303–312. https://doi.org/10.1007/s00704-020-03089-4.

Han, J., Kamber, M., 2012. Data mining: concepts and techniques, 3rd ed. Elsevier, Burlington, MA.

Hargreaves, G.H., Allen, R.G., 2003. History and evaluation of hargreaves evapotranspiration equation. J. Irrig. Drain. Eng. 129, 53–63. https://doi.org/10.1061/(asce)0733-9437(2003)129:1(53).

Hargreaves, G.H., Samani, Z.A., 1982. Estimating potential evapotranspiration. J. Irrig. Drain. Div. 108, 225–230. https://doi.org/10.1061/JRCEA4.0001390.

Ippolito, M., De Caro, D., Cannarozzo, M., Provenzano, G., Ciraolo, G., 2024. Evaluation of daily crop reference evapotranspiration and sensitivity analysis of FAO Penman-Monteith equation using ERA5-Land reanalysis database in Sicily Italy. Agricultural Water Management 295, 108732. https://doi.org/10.1016/j.agwat.2024.108732.

Jia, H., Ding, S., Xu, X., Nie, R., 2014. The latest research progress on spectral clustering. Neural Comput & Applic 24, 1477–1486. https://doi.org/10.1007/s00521-013-1439-2.

Kaufman, L., Rousseeuw, P., 2005. Finding groups in data: an introduction to cluster analysis. Wiley series in probability and mathematical statistics, Wiley, Hoboken, N.J.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Kousari, M.R., Asadi Zarch, M.A., Ahani, H., Hakimelahi, H., 2013. A survey of temporal and spatial reference crop evapotranspiration trends in Iran from 1960 to 2005. Clim. Change 120, 277–298. https://doi.org/10.1007/s10584-013-0821-5.

Koutsoyiannis, D., 2025. When are models useful? revisiting the quantification of reality checks. Water 17, 264. https://doi.org/10.3390/w17020264.

Malamos, N., Tsirogiannis, I.L., Tegos, A., Efstratiadis, A., Koutsoyiannis, D., 2017. Spatial interpolation of potential evapotranspiration for precision irrigation purposes. European Water 303–309.

Mamassis, N., Mazi, K., Dimitriou, E., Kalogeras, D., Malamos, N., Lykoudis, S., Koukouvinos, A., Tsirogiannis, I., Papageorgaki, I., Papadopoulos, A., Panagopoulos, Y., Koutsoyiannis, D., Christofides, A., Efstratiadis, A., Vitantzakis, G., Kappos, N., Katsanos, D., Psiloglou, B., Rozos, E., Kopania, T., Koletsis, I., Koussis, A. D., 2021. OpenHi.net: a synergistically built, national-scale infrastructure for monitoring the surface waters of Greece. Water 13, 2779. https://doi.org/10.3390/w13192779.

McMahon, T.A., Finlayson, B.L., Peel, M.C., 2016. Historical developments of models for estimating evaporation using standard meteorological data. WIREs Water 3, 788–818. https://doi.org/10.1002/wat2.1172.

Monteith, J.L., 1981. Evaporation and surface temperature. Q. J. R. Meteorolog. Soc. 107, 1–27. https://doi.org/10.1002/qj.49710745102.

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

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proc. R. Soc. Lond.Series A. Mathematical and Physical Sciences 193, 120–145. https://doi.org/10.1098/rspa.1948.0037.

Pereira, L.S., Paredes, P., Jovanovic, N., 2020. Soil water balance models for determining crop water and irrigation requirements and irrigation scheduling focusing on the FAO56 method and the dual Kc approach. Agric Water Manag 241, 106357. https://doi.org/10.1016/j.agwat.2020.106357.

Pimentel, R., Arheimer, B., Crochemore, L., Andersson, J.C.M., Pechlivanidis, I.G., Gustafsson, D., 2023. Which Potential Evapotranspiration Formula to Use in Hydrological Modeling World-Wide? Water Resour. Res. 59, e2022WR033447. https://doi.org/10.1029/2022WR033447.

Ritchie, J.T., 1972. Model for predicting evaporation from a row crop with incomplete cover. Water Resour. Res. 8, 1204–1213. https://doi.org/10.1029/WR008i005p01204.

Tegos, A., Efstratiadis, A., Koutsoyiannis, D., 2013. A Parametric Model for Potential Evapotranspiration Estimation Based on a Simplified Formulation of the Penman-Monteith Equation, in: Alexandris, S. (Ed.), Evapotranspiration - An Overview. InTech, pp. 143–165. https://doi.org/10.5772/52927.

Tegos, A., Efstratiadis, A., Malamos, N., Mamassis, N., Koutsoyiannis, D., 2015a. Evaluation of a parametric approach for estimating potential evapotranspiration across different climates. Agric. Agric. Sci. Procedia 4, 2–9. https://doi.org/10.1016/j.aaspro.2015.03.002.

Tegos, A., Malamos, N., Efstratiadis, A., Tsoukalas, I., Karanasios, A., Koutsoyiannis, D., 2017. Parametric modelling of potential evapotranspiration: a global survey. Water 9, 795. https://doi.org/10.3390/w9100795.

Tegos, A., Malamos, N., Koutsoyiannis, D., 2015b. A parsimonious regional parametric evapotranspiration model based on a simplification of the Penman–Monteith formula. J. Hydrol. 524, 708–717. https://doi.org/10.1016/j.jhydrol.2015.03.024.

Tseng, W.-K., Lee, H.-S., 2007. The vector function for distance travelled in great circle navigation. J. Navigation 60, 158–164. https://doi.org/10.1017/S0373463307214122.

Umutoni, L., Samadi, V., 2024. Application of machine learning approaches in supporting irrigation decision making: a review. Agric Water Manag 294, 108710. https://doi.org/10.1016/j.agwat.2024.108710.

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat Comput 17, 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Webb C.P., 2010. Bureau of Meteorology Reference Evapotranspiration Calculations. Climate Services Centre, Queensland Regional Office, Queensland.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213.