The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2026

Volume 40, Pages 360-373

ICBAST 2026: International Conference on Basic Sciences and Technology

Comparative Analysis of Clustering Algorithms for Meteorological Regime Characterization Over Türkiye Using ERA5 Reanalysis Data

Ali Utku Akar
Konya Technical University

Cevat Inal
Konya Technical University

Abstract: Precise analysis and region-specific classification of meteorological parameters constitute a critical prerequisite for advanced scientific processes, ranging from artificial intelligence modeling to complex atmospheric research. The multidimensional and inherently volatile nature of these parameters often hinders the extraction of latent patterns through conventional statistical methods. At this juncture, clustering analysis facilitates the derivation of optimized analytical domains by automatically classifying data points with shared climatological signatures. This study adopts a comparative perspective by evaluating data-driven (K-means and K-medoids), spatial-based (Spatially Constrained Multivariate Clustering: SCMC), and density-based (DBSCAN) clustering approaches to reveal regional variability and spatial dependency across Türkiye. The primary motivation of this research is to determine which approach reflects the heterogeneous structure with high representational power across diverse topographical settings, thereby establishing an optimal clustering strategy for atmospheric modeling. This study aims to characterize and regionally decompose Türkiye's meteorological profile by leveraging high-resolution European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis parameters, including temperature, pressure, relative humidity, instantaneous moisture flux, geopotential, 2m temperature and dewpoint temperature. Methodologically, the ERA5 were downscaled to the geographical coordinates of meteorological stations via point-based extraction and validated against in-situ observations provided by the Turkish State Meteorological Service. Performance analyses indicate that the SCMC approach achieved the highest precision in delineating meteorological similarities and disparities. The comparative success ranking was established as SCMC>DBSCAN>K-medoids>K-means. Findings demonstrate that while K-means and K-medoids are effective in identifying broad climatic zones, the SCMC algorithm generates significantly more consistent and spatially continuous clusters. Furthermore, DBSCAN successfully isolated microclimatic regions with extreme values, such as the Eastern Black Sea, from the primary clusters. Consequently, this research demonstrates that integrating high-resolution reanalysis data with advanced spatial clustering techniques provides a more precise and dynamic decision-support mechanism for regional meteorological analysis and environmental planning compared to conventional methodologies.

Keywords: ERA5 reanalysis data, Clustering analysis, Meteorological research

Introduction

The global climate system is a dynamic structure shaped by the complex interactions of atmospheric circulation, ocean currents, and thermodynamic processes (Bridgman & Oliver, 2014). Over the last seven decades, accelerating anthropogenic impacts have disrupted this natural balance, making "global warming" and the resulting "climate change" the most significant environmental threat facing humanity (Khan, 2024). Understanding the local and regional effects of climate change requires not only tracking temperature increases but also clearly partitioning how meteorological regimes evolve both temporally and spatially (Xie et al., 2015).

The high dimensionality and non-linear nature of meteorological data can render traditional statistical methods insufficient in explaining such complexity (Singh et al., 2024). At this juncture, clustering analysis, a prominent machine learning and data mining technique, emerges as a critical tool for identifying hidden patterns within vast datasets and defining sub-regions or regimes with similar characteristic features (Morawiec et al., 2026). Observations located within the same cluster exhibit a higher degree of similarity to one another based on a specific criterion compared to observations in other clusters. The fundamental principle is to establish a structure where intra-group variation is minimized while inter-group variation is maximized. The origins of clustering analysis date back to the anthropological studies of Driver and Kroeber (1932) and were later adapted to psychology. Cattell (1943) utilized this method extensively for trait theory classifications in personality psychology. As a core step of exploratory data analysis, this method is frequently employed in statistical data processing (Everitt et al., 2011). It holds a significant place in various scientific and engineering fields, such as medicine and health sciences, economics and finance, information technology, telecommunications, and machine learning.

Clustering analysis is generally implemented in four stages: constructing the data matrix, calculating similarity or distance matrices, determining the clustering methods to be employed, and interpreting the obtained results. The construction of the data matrix constitutes the initial phase of the analysis. Primarily, observation values for n units across p variables are obtained from populations for which no definitive information regarding natural classifications exists (Anderberg, 2014). During the selection of variables for the research, the theoretical and practical aspects of the subject must be considered, and the current literature should be reviewed. Like other multivariate statistical methods, clustering analysis does not possess the capability to rectify errors made during the variable selection stage. Furthermore, extreme outliers—observations that deviate significantly from the general trend—must be identified at this stage. In the second phase of the analysis, similarity or distance matrices are calculated (Everitt et al., 2011). In the third stage, the most appropriate method among existing techniques is selected, and clusters are generated. The fourth and final stage involves the interpretation of the results. In determining the correct number of clusters and evaluating the suitability of the structure, clusters consisting of only one or two observations should be viewed with caution. Care should be taken to ensure that the variables used in clustering exhibit significant differences between clusters. Additionally, it is essential that the results are theoretically valid and consistent with the existing literature (Kaufman & Rousseeuw, 2009). The methodology of clustering analysis offers a wide variety of techniques depending on the objective of the study and the structure of the data. Traditionally, these are categorized into two main groups: hierarchical techniques, based on the presentation of results, and non-hierarchical techniques, which partition the dataset according to a pre-determined number of clusters (Silahtaroğlu, 2013). Recently, location-based techniques that establish relationships between data attributes and geographic coordinates have gained critical importance in both academic and practical applications. Through statistics such as spatial autocorrelation, these approaches enable the identification of regional differences by determining the clustering patterns of areas with high, low, or outlier values for any given variable in the dataset (Şişman, 2021).

Clearly revealing spatial variability is of vital importance for developing “regional and specific” strategies rather than “one-size-fits-all” solutions in the fight against climate change. For instance, while drought regimes become dominant in one region, the intensification of extreme precipitation regimes in a neighboring area necessitates distinct adaptation policies for each. This study aims to emphasize the methodological power of clustering analysis in partitioning meteorological regimes and to discuss the role of the precise spatial data provided by these analyses in creating climate-resilient cities and sustainable ecosystems. By evaluating data-driven (K-means and K-medoids), spatial-based (Spatially Constrained Multivariate Clustering: SCMC), and density-based (Density-Based Spatial Clustering of Applications with Noise: DBSCAN) clustering approaches, this study aimed to reveal regional variability and spatial dependency across Türkiye from a comparative perspective. In this context, Türkiye’s meteorological profile was characterized and regionally decomposed by leveraging high-resolution ERA5 reanalysis parameters from the European Centre for Medium-Range Weather Forecasts (ECMWF), including temperature, pressure, relative humidity, instantaneous moisture flux, geopotential, 2m temperature, and dewpoint temperature.

Method

Study Area

The study area encompasses a wide geography across Türkiye, selected to represent the region’s diverse geoclimatic conditions (Figure 1). Türkiye exhibits significant climatic complexity due to its unique geographical position as a bridge between Europe and Asia, surrounded by three distinct seas. This diversity

ranges from temperate Mediterranean and Aegean maritime climates—characterized by high humidity and substantial precipitable water vapor—to semi-arid and continental conditions in Central and Eastern Anatolia, where pronounced temperature gradients and strong seasonal variability are observed. Furthermore, the complex topography, including the North Anatolian and Taurus mountains ranges, generates distinct microclimates and pronounced orographic effects that significantly influence atmospheric delay and signal propagation. Overall, this multifaceted meteorological environment provides an ideal natural laboratory for evaluating the performance and robustness of cluster-based models under varying atmospheric regimes and pressure–temperature profiles.

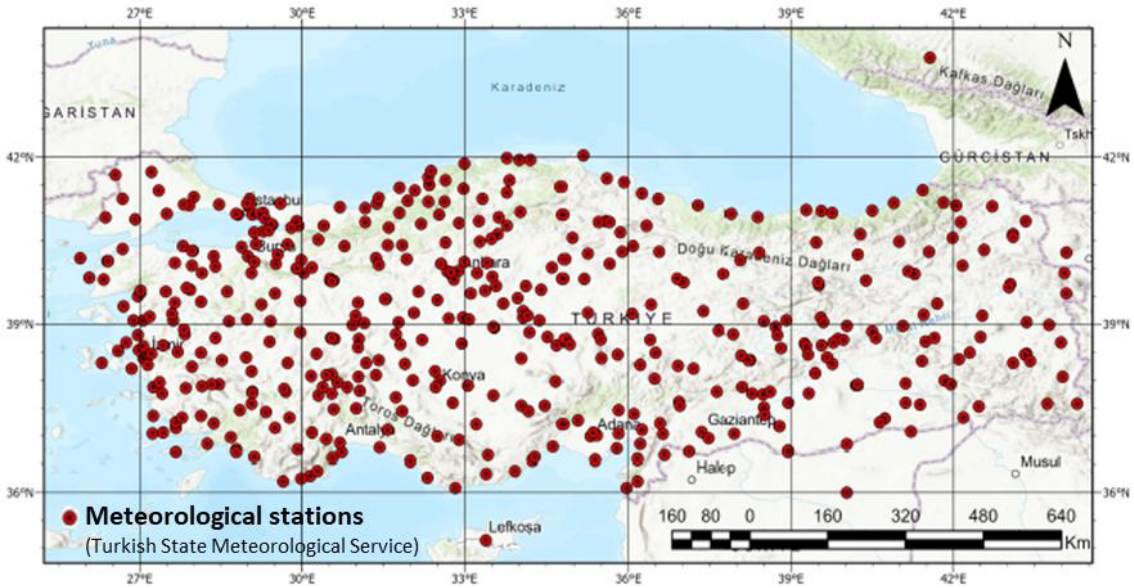


Figure 1. Study area and spatial distribution of meteorological stations

Data Acquisition and Processing

ERA5 reanalysis data is used in this study. Meteorological parameters extracted from ERA5 serve as the basis for the clustering analysis, enabling the identification of spatial regimes within the study area. Figure 2 shows the monthly mean maps of meteorological parameters, covering the five-year period from 2020 to 2024. Here, p*ressure (P)* represents the pressure force per unit area exerted on land, ocean, or inland water surfaces. *2m temperature (T2m)* denotes the air temperature at 2 m above the surface, while *2m dewpoint temperature (DT2m)* represents the dew point temperature at the same height and serves as an indicator of atmospheric humidity. *Geopotential (GEO)* refers to the gravitational potential energy per unit mass at a specific location relative to mean sea level, representing the work required to raise a unit mass against gravity. *Temperature (T)* describes atmospheric temperature available at multiple pressure levels along the vertical profile. *Relative humidity (RH)* quantifies the water vapor mass per kilogram of moist air, where moist air consists of dry air, water vapor, cloud liquid water, cloud ice, rain, and falling snow. Finally, *instantaneous moisture flux (IMF)* is defined as the net moisture exchange rate between the surface and the atmosphere, resulting from evapotranspiration and condensation processes over a given timeframe.

K-Means Clustering

K-means clustering is one of the most fundamental and widely used methods in unsupervised learning. The primary objective of this algorithm is to partition a dataset consisting of $\boldsymbol{n}$ data points into a pre-determined number of $\boldsymbol{k}$ clusters, thereby maximizing intra-cluster similarity and minimizing inter-cluster differences. As a centroid-based algorithm, each cluster is defined by a cluster center ($\boldsymbol{\mu_j}$) representing the average position of the points within that cluster (Aldino et al., 2021). The process operates through an iterative cycle starting from $\boldsymbol{k}$ randomly selected initial centers and consists of two main steps: assignment and update. During the assignment step, each data point $\boldsymbol{x_i}$ is assigned to the nearest center $\boldsymbol{\mu_j}$, typically using Euclidean distance as the proximity metric. In the update step, the new center of each cluster is recalculated by taking the arithmetic mean of all data points assigned to that cluster, which ensures the repositioning of centers to provide the minimum Within-Cluster Sum of Squares (WCSS) (Kriegel et al., 2017).

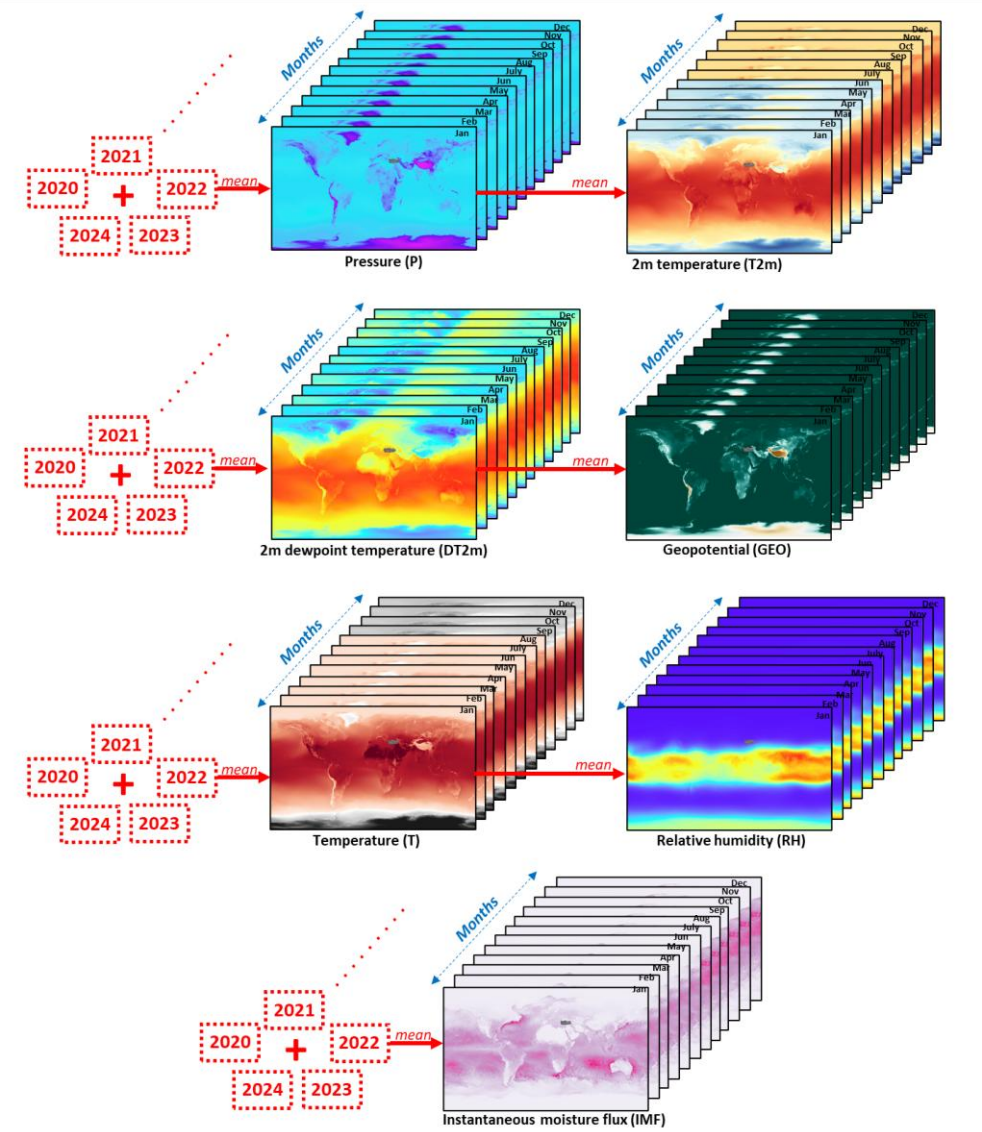


Figure 2. Monthly means spatial distributions derived from meteorological parameters extracted from the ERA5 reanalysis dataset

One of the greatest challenges of the algorithm is determining the optimum number of clusters (***k***). To make this selection and evaluate clustering quality, several methods are utilized, including the Elbow method, Silhouette analysis, Gap statistics, the Davies–Bouldin index, and the Calinski-Harabasz index. Furthermore, the Rand index and the Adjusted Rand index, which are correct for chance groupings, are employed to measure the agreement between two different clustering results.

K-Medoids Clustering

K-medoids is an unsupervised learning and data clustering method similar to the K-means approach. However, unlike K-means, the K-medoids algorithm selects cluster centers from actual observations within the cluster (medoids) rather than using averages. This feature makes K-medoids particularly robust against outliers and noise (Lund et al., 2021). The fundamental basis of K-medoids clustering is to find ***k*** clusters among ***n*** observations by first identifying a random representative observation for each cluster. Each remaining observation is then assigned to the medoid to which it is most similar. Instead of taking the average value of the observations in each cluster, K-medoids utilizes these representative observations as reference points (Kaur et al., 2014). The approach takes the input parameter ***k***, which represents the number of clusters to be partitioned among the set of ***n*** observations, from the user. It operates based on the principle of minimizing the sum of dissimilarities between each observation and its corresponding reference point. This process is executed using the following equation (1):

$$\sum_{i=1}^{k} \sum_{x_j \epsilon S_i} d(x_j, m_i) \quad (1)$$

Spatial-Based Clustering

While most clustering methods consider the similarities between observations or variables in a dataset, they often neglect location information, leading to clusters formed without regard for geographic or spatial relationships. To address this limitation, SCMC has been developed as an approach that incorporates both spatial and non-spatial attributes into the analysis. Based on the Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm, SCMC is considered an unsupervised machine learning technique because it does not require pre-classified features or training data to identify natural patterns in the data. The process assumes a network structure of nodes and edges, aiming to connect nodes with the lowest possible cost through a Minimum Spanning Tree (MST) algorithm. To restrict cluster membership to contiguous or nearby features, the method first establishes a connectivity graph representing spatial relationships. From this graph, an MST is designed to summarize both spatial relationships and data similarity, where each edge weight is proportional to the similarity of the connected objects. While various algorithms like Kruskal (1956) or Boruvka can be used, the approach proposed by Assunção (2006) typically utilizes the Prim algorithm. Once the MST is constructed, the tree is pruned by removing specific branches to create separate clusters. This pruning process selects edges that minimize internal cluster diversity while avoiding single-feature clusters. Although the algorithm optimizes cluster similarity at each iteration, there is no guarantee that the final result will be globally optimal (Pettie & Ramachandran, 2002).

Density-Based Clustering

DBSCAN is a powerful unsupervised machine learning algorithm that focuses on the density distribution of points within a dataset and does not require the number of clusters to be predefined. Unlike traditional centroid-based algorithms, DBSCAN categorizes data into "core points", "border points" and "noise" (outliers), allowing it to successfully identify clusters of arbitrary shapes. The performance of the algorithm primarily depends on two hyper-parameters: the radius distance (ϵ), which defines a point's neighborhood, and the minimum number of points (***MinPts***) required to form a cluster within that radius. During the execution of the algorithm, if the number of points within the **ϵ-neighborhood** of a point is equal to or greater than ***MinPts***, that point is defined as a core point and a new cluster is initiated. All core points that are density-reachable are merged into the same cluster, while low-density points that cannot be assigned to any dense region are labeled as "noise" and excluded from the clusters. This feature makes DBSCAN particularly effective in complex datasets containing outliers and non-linear spatial structures.

Results and Discussion

Clustering algorithms with different functional principles were employed to group meteorological data and identify regional similarities/differences. Following the acquisition of station-scale datasets from ERA5, K-means, K-medoids, SCMC and DBSCAN were applied. Seven parameters related to climatic characteristics (T, GEO, P, IMF, RH, T2m, and DT2m) were utilized as input variables for the clustering analysis and subsequently evaluated statistically. All analyses were conducted using Python (v3.12.7, Spyder 5), IBM SPSS Statistics, and ArcGIS Pro.

K-Means Clustering Results

To determine the optimal number of clusters (***k***) for K-means analysis, the Elbow Method was employed by calculating the WCSS for each ***k*** value. The analysis identified two potential inflection points at ***k***=4 and ***k***=10. While ***k***=4 provided a simplified model representing the fundamental structure of the dataset, ***k***=10 allowed for a more granular decomposition of sub-segments and small-scale variations. Due to the complex and heterogeneous nature of the meteorological data, ***k***=10 was selected as the final cluster count to achieve higher precision. Following the optimization of all parameters, the final clustering was executed to delineate regional patterns across Türkiye (Figure 3).

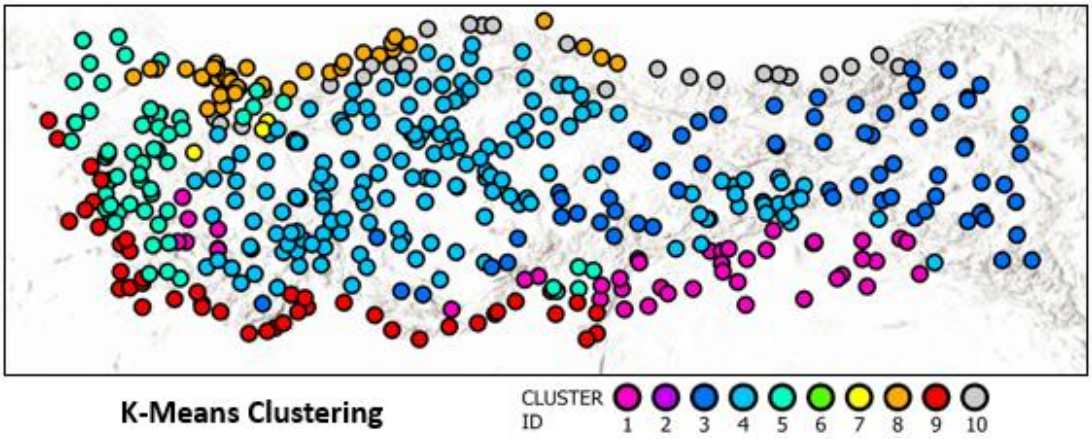


Figure 3. Cluster groups formed as a result of the K-means application

While the clusters diverge at specific locations, the spatial separation may not appear fully distinct. This is primarily because the method is applied directly to the numerical dataset; although stations with similar meteorological characteristics are grouped together, their explicit spatial weights and geographical relationships are not fully integrated into the algorithm. As seen in Figure 3, cluster distributions emerge from the coastal regions toward the interior.

To identify cluster groups with varying meteorological values in Türkiye through statistical analysis, box plots were generated. The relationship between monthly meteorological data belonging to these clusters was evaluated via descriptive statistics. In the box plots, colors represent the respective cluster groups (Figure 4). Minimum, maximum, mean, quartiles, and outliers were analyzed, and overlapping box plots were identified as similar clusters. This process was performed separately for each month and parameter. The meteorological parameters were processed using IBM SPSS Statistics to determine their significance levels, which are indicated by the below Table 1.

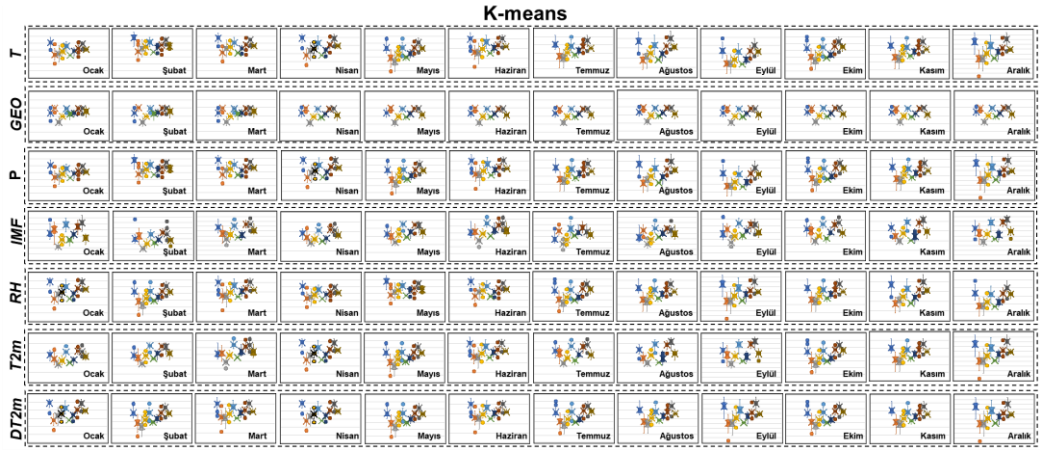


Figure 4. Box plots created based on monthly parameters in cluster groups determined by K-means

Table 1. Identifying clusters that are similar to each other among the clusters obtained from K-means

Cluster	1	3	4	5	7	8	9	10
T		■		■		■		■
GEO				■		■	■	
P				■		■	■	
IMF		■			■			
RH			■		■			
T2m	■					■	■	
DT2m			■	■				
	1	2	2	**4**	2	**4**	3	1

Based on the black-filled cells in Table 1, it is observed that Clusters 5 and 8 exhibit the highest frequency of shared meteorological parameters across all evaluations. This indicates that these two identified cluster groups possess closely aligned data characteristics. Conversely, no significant similarities or correlations were found among the parameters of Clusters 1, 3, 4, 7 and 10.

3K-Medoids Clustering Results

The Silhouette Method was employed to determine the optimal number of clusters (***k***) for the K-medoids algorithm. As the silhouette coefficient evaluates both the cohesion of an observation within its own cluster and its separation from neighboring clusters, it provides a robust validity metric aligned with the K-medoids approach. The most meaningful cluster count for the dataset was identified by comparing the average silhouette scores calculated for various ***k*** values.

Analysis of the silhouette scores revealed that the linear upward trend in the average score stabilized at ten clusters; consequently, ***k***=10 was selected as the optimal value. Generally, an average silhouette width exceeding 0.7 is considered indicative of a "strong" clustering structure, while values between 0.5 and 0.7 are deemed "reasonable", and those between 0.25 and 0.5 are classified as "weak". In the scenario where ***k***=10, the silhouette width exceeded the 0.7 threshold, confirming that the K-medoids clustering provides a highly robust and well-defined representation of the data.

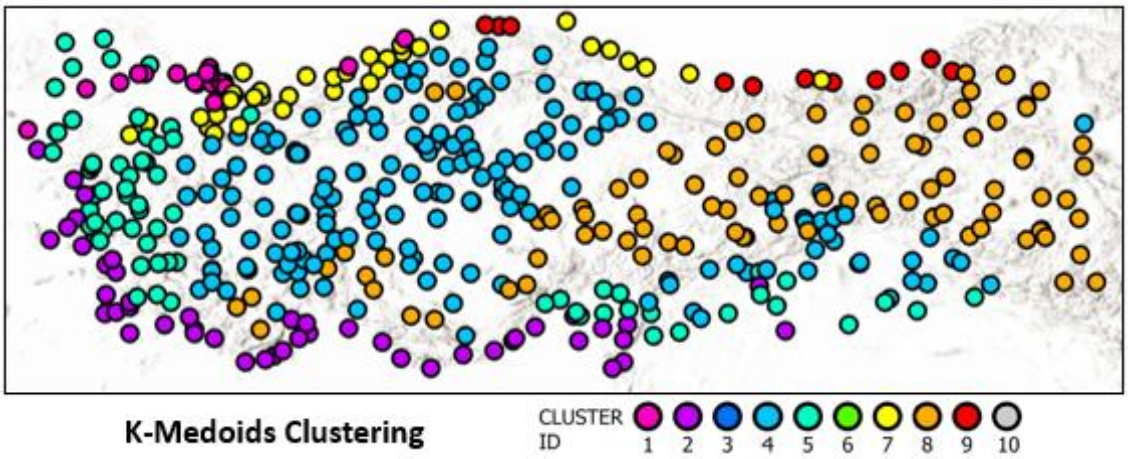


Figure 5. Cluster groups formed as a result of the K-medoids application

Figure 5 illustrates the clustering map resulting from the K-medoids application. While Clusters 4 and 8 correspond to the regions of Türkiye characterized by a terrestrial climate, Clusters 1, 2, 5, 7, and 9 exhibit a distinct separation toward the coastal areas, driven by the influence of RH and the IMF. The box plots generated for the statistical identification of cluster groups with similar or divergent meteorological values in Türkiye are presented in Figure 6. The corresponding data derived from this analysis are summarized in Table 2.

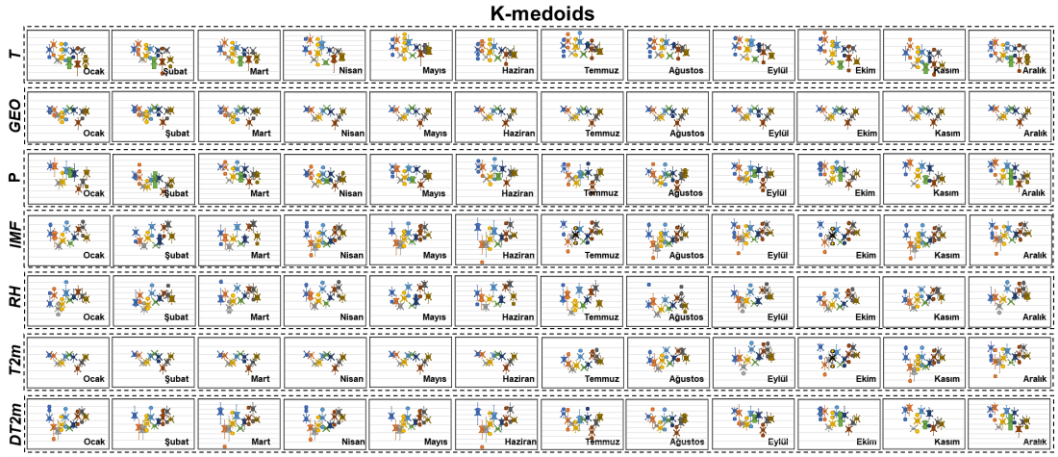
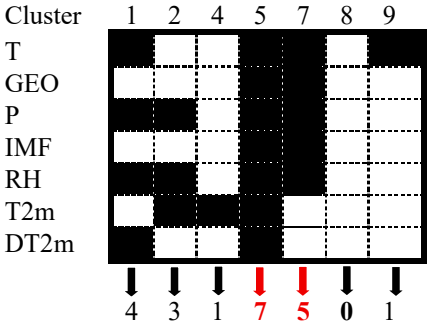


Figure 6. Box plots created based on monthly parameters in cluster groups determined by K-medoids

Table 2. Identifying clusters that are similar to each other among the clusters obtained from K-medoids



Based on the results presented in Table 2, clusters identified as meteorologically similar in K-means were observed to diverge in the K-medoids application. The fundamental difference between these two algorithms stems from their sensitivity to outliers: while K-means determines cluster centers based on the mean value, K-medoids selects an actual data point within the cluster as the center (the medoid), making it more robust against outliers. Specifically, while Clusters 5 and 8 were identified as the most similar in K-means, Clusters 5 and 7 emerged as the groups containing the most common parameters in the K-medoids analysis.

SCMC Results

Clusters were generated via ArcGIS Pro under spatial constraints, accounting for both geographical location and multivariate characteristics. In contrast to the previously mentioned clustering methods, the initial number of clusters for SCMC was set to 10; subsequently, optimization parameters—such as cluster size constraints (number of clusters), spatial constraints (trimmed delaunay triangulation), and membership probabilities—were utilized to achieve the most effective differentiation among these ten clusters.

The clustering map derived from the SCMC analysis is presented in Figure 7. Distinct spatial boundaries are observed between the clusters in Türkiye, particularly for Clusters 2, 4, 5, 7, 8, 9, and 10. This indicates that climate, topography, and spatial factors exert a strong influence on the formation of these clusters.

It is evident that SCMC serves as a significant method for clustering geographical and climatic parameters. Furthermore, it can be utilized as a strategic guide in geography-based decision-making processes; however, the cluster parameters must be statistically examined. To this end, an approach was adopted where spatial and non-spatial characteristics were handled holistically. Based on the box plots presented in Figure 8, similar clusters were identified and marked in Table 3.

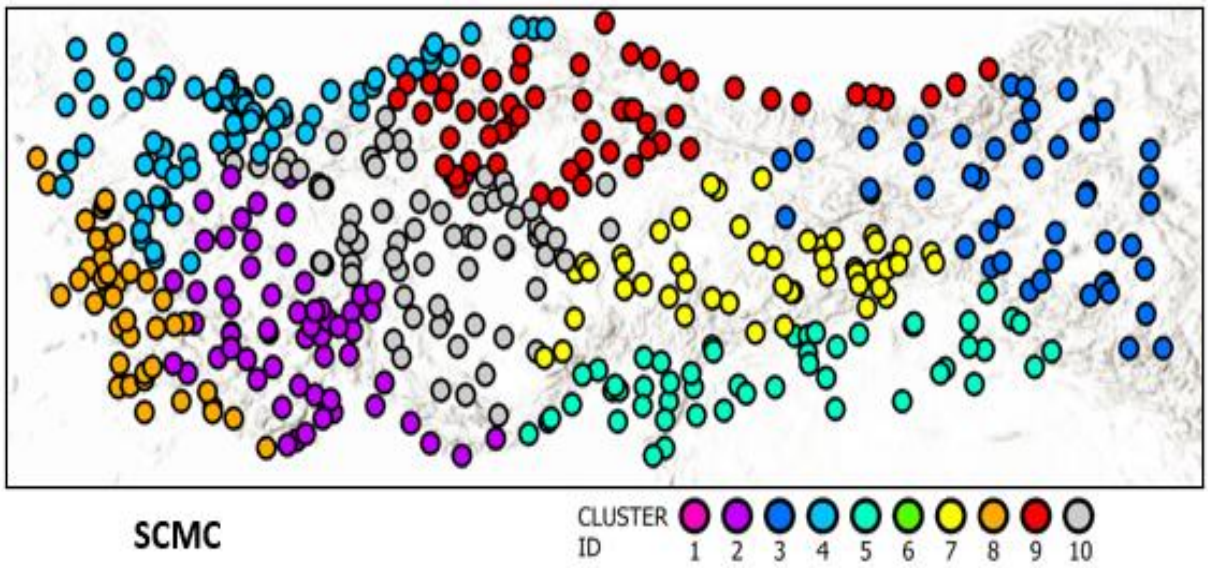


Figure 7. Cluster groups formed as a result of the SCMC

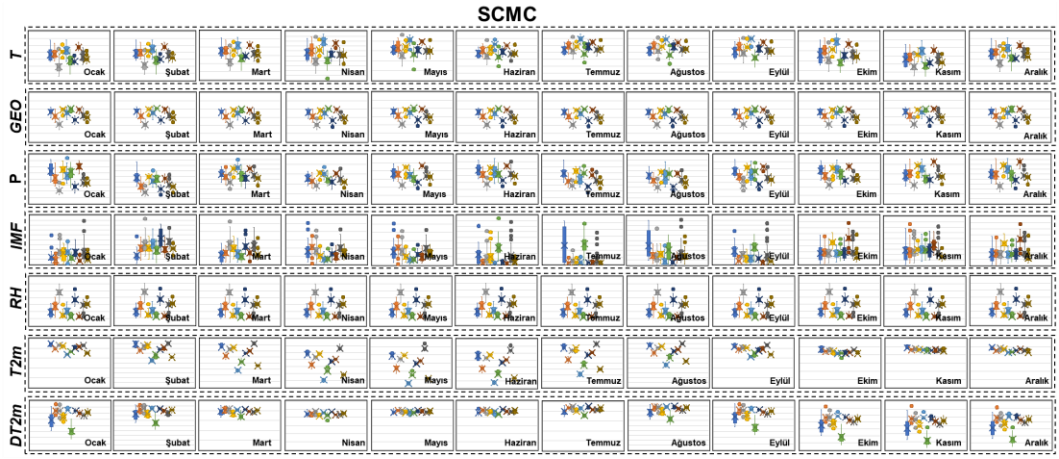
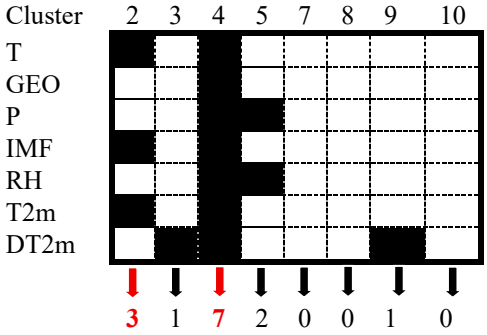


Figure 8. Box plots created based on monthly parameters in cluster groups determined by SCMC algorithm

Table 3. Identifying clusters that are similar to each other among the clusters obtained from SCMC



Based on the results derived from Table 3, it is observed that Clusters 2 and 4 are characteristically very similar. Evaluating the T, GEO, P, IMF, RH, T2m, and DT2m parameters on a monthly basis reveals that these variables overlap for Clusters 2 and 4 across almost all months.

DBSCAN Clustering Results

In this study, the maximum search distance (ϵ) and the minimum number of features required per cluster were determined through an iterative optimization process. Cluster memberships were identified based on these density criteria. When the minimum feature requirement per cluster is low (15), the differentiation of meteorological data is poorer, resulting in fewer clusters. In other words, due to the lenient density criteria, the algorithm is likely to merge most stations under a single cluster. Conversely, a higher minimum feature count (45) implies a requirement for stronger similarity among station data, leading to more refined cluster separation. Consequently, a scenario with a minimum point count of "45", a cluster count of "8", and a search distance of "150" was selected.

The clustering map generated using the selected parameters in the DBSCAN algorithm is presented in Figure 9. Unlike other clustering methods, noisy data were excluded from the map in the DBSCAN analysis. These noisy data represent stations that could not be assigned to any specific cluster due to their failure to meet the required DBSCAN criteria, specifically the search distance (ϵ) and the minimum number of features per cluster (***MinPts***). Based on the box plots presented in Figure 10, similar clusters were identified and marked in Table 4.

The comparative analysis of meteorological parameters across the identified clusters reveals a significant degree of similarity between Cluster 2 and Cluster 4. As illustrated in the parameter intersection matrix, Cluster 4 stands out as the most representative group, exhibiting commonalities across six out of seven evaluated parameters. Cluster 2 follows this trend by sharing three key parameters (T, IMF, and T2m) with the overall meteorological profile. The high frequency of black-filled cells for Cluster 4, particularly when compared with the overlap in Cluster 2, suggests that these two groups possess closely aligned climatic characteristics. While other clusters (such as 7 and 8) show no significant parameter intersections, the strong convergence in Clusters 2 and 4 indicates that the stations within these groups respond similarly to the multivariate atmospheric variables.

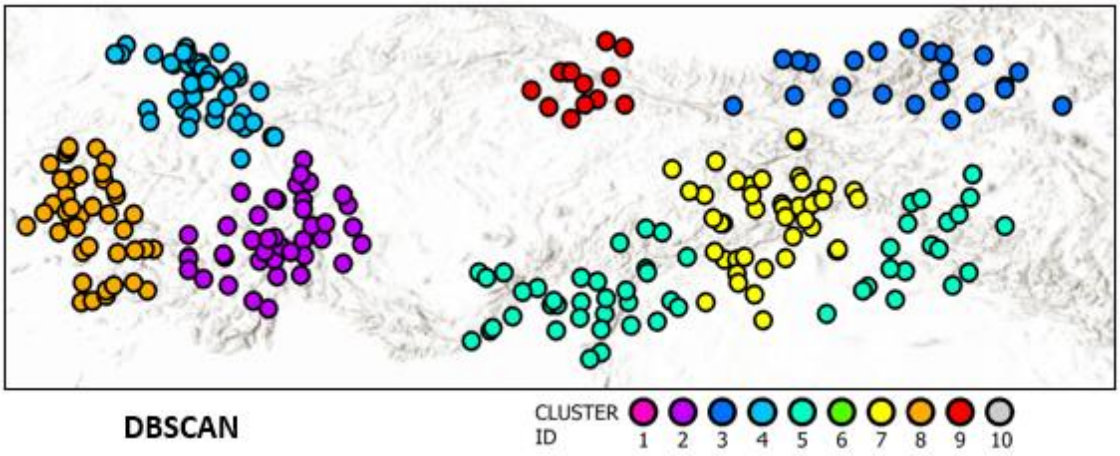


Figure 9. Cluster groups formed as a result of the DBSCAN

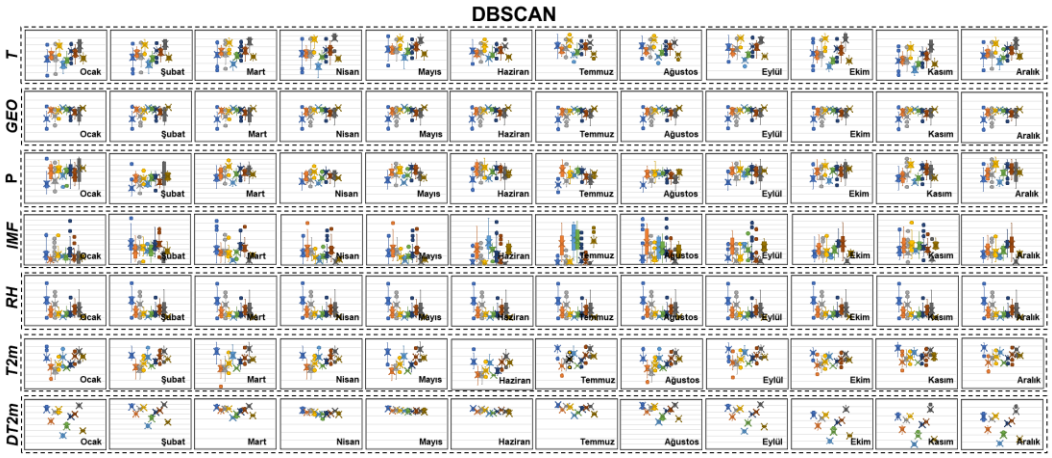
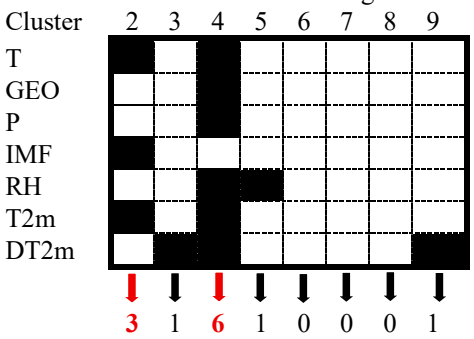


Figure 10. Box plots created based on monthly parameters in cluster groups determined by DBSCAN

Table 4. Identifying clusters that are similar to each other among the clusters obtained from DBSCAN



Overall Assessment of Clustering Algorithm Results

The similarities and divergences between cluster groups have been statistically established using K-means, K-medoids, SCMC, and DBSCAN approaches. However, it has not yet been determined which method most accurately reflects the characteristic meteorological regions. Specifically, the question of which algorithm best captures meteorological diversity remains to be addressed. To identify the most successful method in this regard, the following evaluative step was implemented.

A "cluster error" was assigned to each cluster across all clustering methods. First, the annual mean values were calculated by averaging the twelve-month meteorological data for each station within its respective cluster. This procedure was performed for all "**n**" stations in each cluster. Subsequently, the overall mean of these "**n**" stations was determined. The cluster score was obtained by calculating the residuals between the individual

meteorological values of each station and the determined cluster mean (Table 5). To eliminate discrepancies arising from the different units of the meteorological parameters, "Min-Max Normalization" was applied to the dataset.

Table 5. Comparison of clustering methods based on calculated cluster errors

Cluster No	K-means	K-medoids	SCMC	DBSCAN
Cluster-1	4.029	3.772	-	-
Cluster-2	-	3.579	3.126	3.508
Cluster-3	4.153	-	2.874	3.180
Cluster-4	4.165	4.100	2.554	2.777
Cluster-5	3.831	3.123	3.007	3.119
Cluster-6	-	-	-	3.112
Cluster-7	4.101	3.117	3.102	3.387
Cluster-8	3.765	3.869	2.664	2.899
Cluster-9	3.648	4.333	3.801	3.823
Cluster-10	4.235	-	3.455	-

Upon evaluating the cluster errors presented in Table 5, it is observed that the SCMC method yields relatively lower cluster errors compared to the other approaches. This indicates that SCMC provides the most robust representation of meteorological regionalization by minimizing intra-cluster variance. The results most closely aligned with those of the SCMC method were achieved by the DBSCAN algorithm, further validating the effectiveness of density-based and spatially constrained techniques in capturing complex climatic patterns. The comparative success of the algorithms is ordered as follows: SCMC > DBSCAN > K-medoids > K-means.

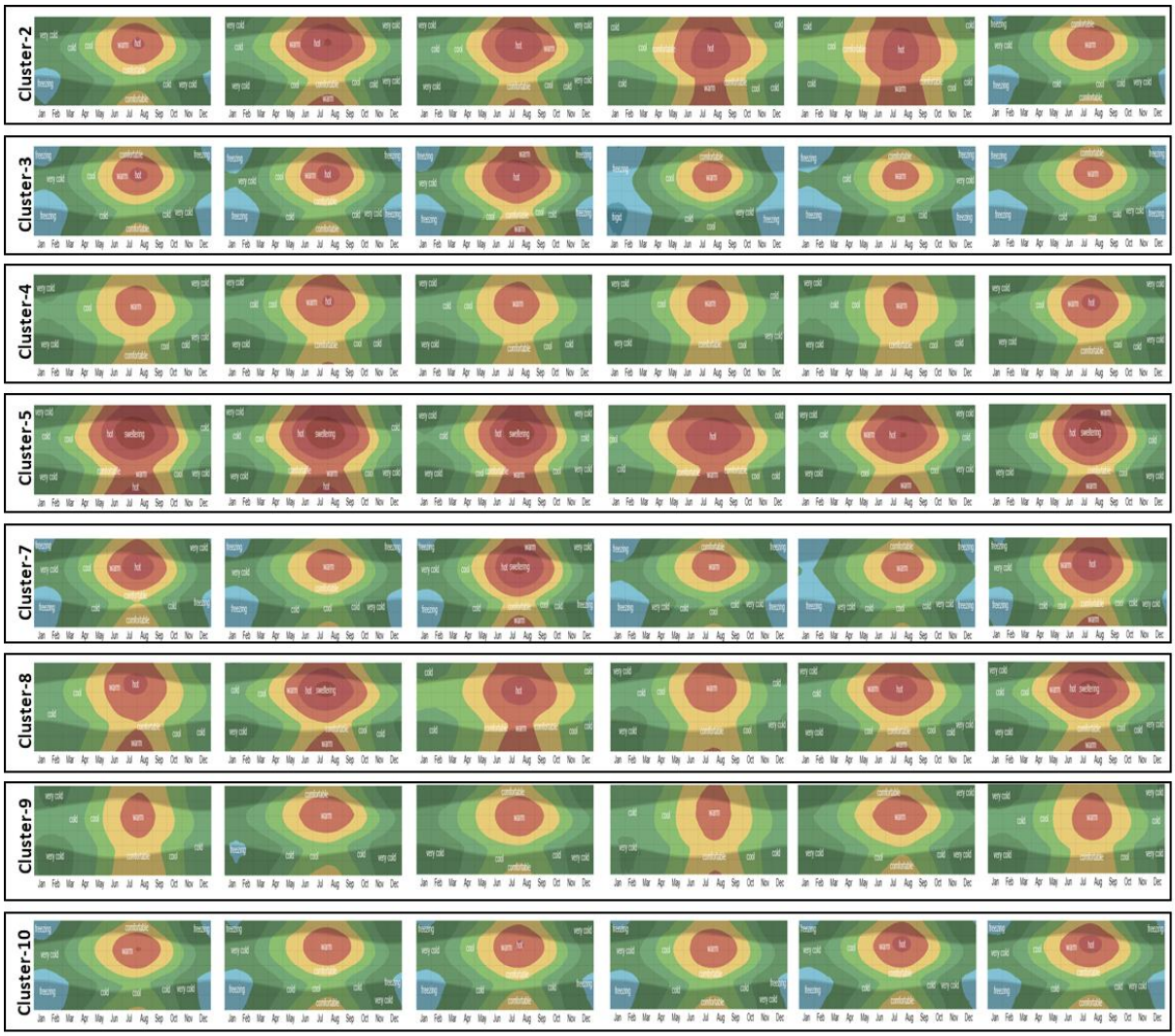


Figure 11. Monthly average temperature heat maps for randomly selected stations (Source: Weather Spark)

This ranking demonstrates that SCMC outperforms traditional methods by effectively integrating geographical constraints with climatic variables. While density-based approaches like DBSCAN follow closely due to their noise-filtering capabilities, partition-based algorithms such as K-means and K-medoids show relatively lower performance in capturing the complex, spatially-dependent structure of meteorological data.

Figure 11 presents the monthly variations of randomly selected stations within the clusters identified by the SCMC algorithm. This visual representation serves as a temporal validation of the clustering success. The analysis reveals two critical findings: high intra-cluster homogeneity and distinct inter-cluster heterogeneity. The temporal profiles in Figure 11 confirm that the SCMC algorithm effectively captures the climatic fingerprints of each region. The stations within the same cluster exhibit nearly identical seasonal patterns, particularly in the peak summer and winter months, indicating a high degree of internal consistency. For instance, the "hot/sweltering" core observed in Cluster-5 is consistently replicated across all sampled stations, while Cluster-3 maintains a dominant "freezing" signature. Furthermore, the inter-cluster variability is visually striking. The differences in the duration and intensity of thermal comfort zones between Cluster-2 and Cluster-10 underscore the high resolution of the SCMC method. The ability of the algorithm to group stations with such synchronized monthly transitions—despite their geographical distance in some cases—proves that the multi-variable and spatial constraint approach provides a superior regionalization compared to traditional methods.

Conclusion

The complex and dynamic nature of the global climate system necessitates advanced analytical approaches to accurately identify regional meteorological regimes. This study demonstrated the critical role of clustering analysis in decomposing the high-dimensional spatial and climatic characteristics of Türkiye, utilizing high-resolution ERA5 reanalysis data. By comparatively evaluating K-means, K-medoids, SCMC, and DBSCAN algorithms, the research highlighted the methodological strengths and limitations of distinct clustering techniques in spatial climatology. The comparative error analysis revealed that traditional partition-based algorithms, such as K-means and K-medoids, exhibit limited capacity in capturing the intricate spatial dependencies of meteorological variables. Because these algorithms primarily focus on the numerical similarities of the data while neglecting geographic proximity, their resulting clusters often lack spatial coherence. In contrast, the SCMC method yielded the lowest cluster errors and emerged as the most successful algorithm. By simultaneously optimizing for both attribute similarity and geographic contiguity, SCMC generated well-defined, robust, and homogenous climate zones. DBSCAN also demonstrated strong performance by effectively filtering out noisy data and identifying high-density regions without requiring a predefined number of clusters, making it the second most effective method in this comparative framework. Ultimately, integrating spatial data with advanced machine learning techniques bridges the gap between raw meteorological observation and actionable climate policy.

Recommendations

Based on the findings of this study, it is highly recommended to prioritize spatially constrained and density-based algorithms, such as SCMC and DBSCAN, over traditional partition-based methods like K-means when analyzing high-dimensional geoclimatic data. The high-resolution clustering maps generated through these advanced techniques should serve as a foundation for policymakers and urban planners to abandon "one-size-fits-all" approaches and instead develop region-specific climate adaptation strategies tailored to localized risks, such as interior drought or coastal extreme precipitation. Furthermore, future research should expand upon this framework by integrating spatiotemporal dynamics to track how these meteorological boundaries shift over time. Incorporating these baseline clusters into predictive machine learning models alongside Global Climate Models (GCMs), and expanding the input variables to include multidisciplinary environmental indices like NDVI and land use changes, will provide a more holistic and predictive understanding of the complex feedback loops between anthropogenic activities and regional climates.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest.

Funding

* The authors received no financial support for the research, authorship, and/or publication of this article.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Basic Sciences and Technology (www.icbast.net) held in Konya/Türkiye on April 02-05, 2026.

* The authors wish to thank the Turkish State Meteorological Service and the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the meteorological data used in this research. This study is also a part of the first author's PhD thesis conducted at the Department of Geomatics Engineering, Faculty of Engineering and Natural Sciences, Konya Technical University.

References

Aldino, A. A., Darwis, D., Prastowo, A. T., & Sujana, C. (2021). Implementation of K-means algorithm for clustering corn planting feasibility area in South Lampung Regency. In *Journal of Physics: Conference Series* (Vol. 1751, No. 1, p. 012038). IOP Publishing.

Anderberg, M. R. (2014). *Cluster analysis for applications* (Vol. 19). Academic Press.

Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science, 20*(7), 797–811.

Bridgman, H. A., & Oliver, J. E. (2014). *The global climate system: Patterns, processes, and teleconnections*. Cambridge University Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*(3), 153–193.

Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology, 31*, 211–256.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). John Wiley & Sons.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kaur, N. K., Kaur, U., & Singh, D. (2014). K-medoids clustering algorithm: A review. *International Journal of Computer Applications Technology, 1*(1), 42–45.

Khan, A. A. (2024). An analytical study on the impact of global warming: Effects on environment. In *Recent trends in commerce, management, accountancy and business economics* (p. 54).

Kriegel, H.-P., Schubert, E., & Zimek, A. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems, 52*(2), 341–378.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society, 7*(1), 48–50.

Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: K-means and k-medoids clustering. *Performance Measurement and Metrics, 22*(3), 161–173.

Morawiec, T., Zareba, M., Danek, T., & Chuchro, M. (2026). Air pollution macro-regions identification using machine learning and spatio-temporal analysis. *PLOS ONE, 21*(1), e0340191.

Pettie, S., & Ramachandran, V. (2002). An optimal minimum spanning tree algorithm. *Journal of the ACM (JACM), 49*(1), 16–34.

Silahtaroğlu, G. (2013). *Veri madenciliği: Kavram ve algoritmaları* (2. baskı). Papatya Yayıncılık.

Singh, G., Sharma, A., Vishwakarma, P., & Gaur, M. (2024). Revealing non-linear trends and patterns in weather data through regression analysis approach. In *Artificial intelligence and information technologies* (pp. 50–56). CRC Press.

Şişman, S. (2021). *CBS tabanlı makine öğrenme teknikleri ile toplu taşınmaz değerlemesi* (Yüksek lisans tezi). Gebze Teknik Üniversitesi, Fen Bilimleri Enstitüsü.

Xie, S. P., Deser, C., Vecchi, G. A., Collins, M., Delworth, T. L., Hall, A., … Watanabe, M. (2015). Towards predictive understanding of regional climate change. *Nature Climate Change, 5*(10), 921–930.

Author(s) Information

Ali Utku Akar
Res. Asst. (PhD)
Konya Technical University
Faculty of Engineering and Natural Sciences,
Department of Geomatics Engineering, Konya, Türkiye
ORCID iD: https://orcid.org/0000-0001-5639-9987
*Contact e-mail: *auakar@ktun.edu.tr*

Cevat Inal
Prof. Dr.
Konya Technical University
Faculty of Engineering and Natural Sciences,
Department of Geomatics Engineering, Konya, Türkiye
ORCID iD: https://orcid.org/0000-0001-8980-2074

*Corresponding author(s) email

To cite this article:

Akar, A. U., & Inal, C. (2026). Comparative analysis of clustering algorithms for meteorological regime characterization over Turkiye using ERA5 reanalysis data. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 40,* 360-373.