Hindawi
Advances in Meteorology
Volume 2022, Article ID 3581299, 14 pages
https://doi.org/10.1155/2022/3581299

*Research Article*

# Data-Driven versus Köppen–Geiger Systems of Climate Classification

**Vajira Lasantha, Taikan Oki, and Daisuke Tokuda**

*Graduate School of Engineering, The University of Tokyo, Tokyo, Japan*

Correspondence should be addressed to Vajira Lasantha; munagama@env.t.u-tokyo.ac.jp

Received 16 March 2022; Accepted 9 August 2022; Published 31 August 2022

Academic Editor: Hisayuki Kubota

Climate zone classification promotes our understanding of the climate and provides a framework for analyzing a range of environmental and socioeconomic data and phenomena. The Köppen–Geiger classification system is the most widely used climate classification scheme. In this study, we compared the climate zones objectively defined using data-driven methods with Köppen–Geiger rule-based classification. Cluster analysis was used to objectively delineate the world's climatic regions. We applied three clustering algorithms—*k*-means, ISODATA, and unsupervised random forest classification—to a dataset comprising 10 climatic variables and elevation; we then compared the obtained results with those from the Köppen–Geiger classification system. Results from both the systems were similar for some climatic regions, especially extreme temperature ones such as the tropics, deserts, and polar regions. Data-driven classification identified novel climatic regions that the Köppen–Geiger classification could not. Refinements to the Köppen–Geiger classification, such as precipitation-based subdivisions to existing Köppen–Geiger climate classes like tropical rainforest (Af) and warm summer continental (Dfb), have been suggested based on clustering results. Climatic regions objectively defined by data-driven methods can further the current understanding of climate divisions. On the other hand, rule-based systems, such as the Köppen–Geiger classification, have an advantage in characterizing individual climates. In conclusion, these two approaches can complement each other to form a more objective climate classification system, wherein finer details can be provided by data-driven classification and supported by the intuitive structure of rule-based classification.

## 1. Introduction

Categorizing regions across the globe based on their climate is beneficial for summarizing climatological data and for explaining and disseminating environmental and sociopolitical data. Climate classification has been used as a basis for regionalization at global and regional scales in various fields, including ecological monitoring [1, 2], hydrology [3, 4], evolutionary anthropology [5], agriculture [6–8], and epidemiology [9, 10]. Many classification systems have been previously reported [11–13]. Among those, the system proposed by Wladimir Köppen [14] is the most famous; it classifies global climate into 30 classes under five main groups, originally intended to be representative of the distribution of five vegetation types described by De Candolle according to the climate zones known to ancient Greeks [15]. The Köppen system describes climate zones based on several climatic variables derived from monthly temperatures and precipitation. Subsequent refinements and modifications were made to the original Köppen system [16, 17]. In this study, we refer to the Köppen–Geiger classification system (hereafter called "KG"), following which world maps of climatic regions were prepared by Grieser et al. [18], Kottek et al. [19], and Peel et al. [20] for 1951–2000; Rubel and Kottek [21] for 1901–2100; and Beck et al. [22] for 1980–2016 and the future.

The KG classification system is a rule-based, top-down approach to climate classification. Classification criteria in KG are tuned for their original purpose of reproducing the distribution of vegetation. Therefore, the climate zones produced by the KG classification system are subjective and limited to predetermined climate types. At the same time, the rule-based nature gives the KG classification system the advantage of being reproducible and intuitive, which is desirable for ease of communication and wide-scale adoption.

KG classification has been critically evaluated by several researchers for its ability to delineate distinct climates around the world. Triantafyllou and Tsonis [23] evaluated the KG classification system by classifying climate stations into Köppen classes on an annual basis and estimating the frequency of changes between major Köppen climate groups. They found that in some regions of the world, KG is unstable to interannual changes in climate. They also reported that KG is either unable or slow to respond to long-term climate change, such as global warming. They suggested that statistical and data-driven approaches, such as factor analysis and cluster analysis, can be the basis for a more objective and robust classification system. Rubel and Kottek [24], in their comments on Köppen's original paper and the review of KG's subsequent developments, remarked that in the past, climate classification was exclusively based on human expertise; however, today, it is supported by various statistical techniques, such as cluster analysis, which can objectively define the global climatic regions.

In this study, we aimed at studying how a data-driven, bottom-up approach to climate classification would produce objectively delineated climate zones and how they could be compared with the outcomes of KG. We present an objective and data-driven classification of the world's climate based on a cluster analysis approach. Previous studies [25–28] have used cluster analysis methods to regionalize the global climate as well as regional climates. DeGaetano [29], Russell and Moore [30], and Unal et al. [31] applied cluster analysis to primary station data to regionalize climates. Fovell and Fovell [32] regionalized the climate in the United States by clustering 344 climate divisions of the National Climatic Data Center (NCDC) dataset. Marston and Ellis [33] and Park et al. [34] applied cluster analysis to gridded climate data to regionalize the climates of the United States and the Korean peninsula, respectively. Hoffman et al. [35] used cluster analysis of the outputs of a general circulation model (GCM) to identify climate regimes and compare different simulation scenarios. Kumar et al. [36] applied a parallel processing implementation of a $k$-means clustering algorithm on large high-dimensional datasets comprising observed, remotely sensed, and simulated data to identify ecoregions in the United States. The choice of input data, their preparation, and the selection of clustering algorithms are important factors that determine the nature of the clusters produced [26].

Our study differs from previous data-driven classifications in the selection and preparation of input data and the setup of clustering methods. We prepared climatic variables similar to those in the KG. While the KG criteria distinguish climate classes using predefined thresholds in the data variables, we intend to reveal natural groupings in the data. Because the input climatic data were selected to be similar, a comparison could be made between the rule-based KG climate classes and data-driven clusters. When setting up the clustering methods, we followed multiple approaches toward two main objectives. One was to minimize subjectivity owing to any prior information given when setting up cluster analyses, such as the number of clusters. For this purpose, we included ISODATA clustering in our study because it does not require prior specification of the number of classes. With the $k$-means and random forest clustering algorithms, which require prior specification of the number of classes, we minimized subjectivity by estimating the optimum number of clusters. Second, to compare data-driven and rule-based classifications, we set up cluster analyses to create the same number of clusters as in the KG.

## 2. Data and Methods

*2.1. Gridded Climatic Data.* Reanalysis data from the Climatic Research Unit gridded Time Series (CRU TS) is a widely used global climate dataset that covers all land areas, except Antarctica [37]. It provides 10 climatic variables at a spatial resolution of 0.5°. For this study, the monthly mean 2 m air temperature and monthly precipitation rate from the CRU TS dataset version 4.05 were used. Because this study was aimed at classifying the present climate, data for the 30-year period from 1991 to 2020 were extracted.

The GMTED2010 global digital elevation model (DEM) was used for obtaining elevation data. It was developed based on data derived from multiple elevation data sources and is available at different resolutions [38].

*2.2. KG Classification.* The KG classification system was adopted following the criteria described by Peel et al. [20], which have also been used by Kriticos et al. [39] and Beck et al. [22]. This classification system has been slightly modified from the original method presented by Köppen [14] and Geiger [16], and the differences have been discussed by Beck et al. [22]. The KG classification criteria are included in the Supplementary Materials (S1). The criteria for classification in KG were defined using 11 climatic variables that were calculated using the monthly precipitation and mean temperature data. These variables were first calculated at a 1-year time resolution. Then, the latest 30-year (1991–2020) averages of these variables were calculated. The updated KG classification of the present climate was prepared by applying the classification criteria to the 30-year averaged variables.

*2.3. Data-Driven Classification.* Ten climatic variables were derived from the precipitation and temperature data for data-driven classification: mean annual temperature ($T_{\text{mean}}$), mean annual precipitation ($P_{\text{year}}$), air temperature of the coldest month in summer ($T_{\text{smin}}$), air temperature of the warmest month in summer ($T_{\text{smax}}$), air temperature of the coldest month in winter ($T_{\text{wmin}}$), air temperature of the warmest month in winter ($T_{\text{wmax}}$), precipitation of the driest month in summer ($P_{\text{sdry}}$), precipitation of the wettest month in summer ($P_{\mathrm{swet}}$), precipitation of the driest month in winter ($P_{\mathrm{wdry}}$), and precipitation of the wettest month in winter ($P_{\mathrm{wwet}}$). The definition of the seasons is consistent with that in the KG system; out of the two 6-month periods (April–September and October–March), the warmer is designated as summer and the colder as winter, at each grid cell. A map of the regions that experienced summer in April–September is presented in Figure 1. This indicates that the two regions are not separated along the equator. Some April–September summer areas were enclaved within the other region. Notably, these areas coincide with the Amazon and Congolian rainforests. All climatic variables were prepared in the form of global 0.5° grids, consistent with that in the original CRU TS data system. The GMTED2010 DEM dataset [38] provides elevation data at various resolutions. The 0.5° resolution elevation data (ELE) were used in this study.

These climatic variables were selected after performing trial clustering exercises. Initial attempts with monthly means of the temperature and precipitation failed to recognize similar climates in different hemispheres because they occurred in different months. The use of seasonal methods allowed us to overcome this drawback. Because the seasons were defined in the same way as in KG, results that are more comparable to KG classes were obtained. However, one may obtain a more objective clustering output with an objective definition of seasons.

*2.4. Principal Component Analysis.* The redundancy of information in the chosen data variables can be a source of bias in the clustering analysis. With regard to their hierarchical clustering analysis of the climate in the United States, Fovell and Fovell [32] discussed the problem of information redundancy. They used principal component analysis (PCA) to reduce the correlation among the data, thereby attempting to reduce redundancy. However, they noted that a certain amount of redundancy can remain even among the principal components (PCs) that are orthogonal to each other.

We employed PCA to reduce redundancy in the data. Furthermore, the reduction in dimensions would help reduce the complexity of the computations. PCA was applied to normalized data variables. Both PCA and normalization have been employed in the cluster analyses of climatic data [Fovell [40], Fovell and Fovell [32], Gómez-Zotano et al. [41], Kozjek et al. [42]]. Netzel and Stepinski [26] highlighted the importance of proper normalization and reported that their modified normalization method for precipitation data performed better than uniform normalization, which tends to produce clusters that are largely influenced by temperature. In this study, we standardized each data variable as a $z$-score.

The number of PCs retained for cluster analysis was determined by inspecting the scree plot. Selecting PCs based on the location of the elbow in a scree plot is an accepted stopping rule in PCA [43]. Three PCs were selected based on the scree plot shown in Figure 2. These represent 90% of the variance in the data. The loadings for the three PCs are listed in Table 1.

*2.5. Cluster Analysis.* Cluster analysis was used in this study as a data-driven approach to delineate climatic regions. Out of the many available clustering algorithms, three were considered in this study.

*2.5.1. k-Means Clustering.* $k$-means clustering is a widely used clustering method in which the $k$ number of partitions is constructed by assigning each observation to the nearest cluster in terms of the distance to the mean of the cluster [44]. The $k$-means clustering algorithm by Macqueen [45] was implemented in the *Cluster* package in *R* version 4.1.1 [46] and used in this study. Euclidean distance was used as the distance function. Three selected PCs of the data variables were used for clustering.

*2.5.2. ISODATA Clustering.* The iterative self-organizing data analysis technique (ISODATA), a partitioning-type clustering method, is a modification of the $k$-means clustering algorithm [47]. Unlike $k$-means clustering, it does not require prior specification of the number of classes. Starting with an initial user-defined number of clusters, the ISODATA algorithm alters the number of clusters by merging, splitting, or deleting clusters based on certain heuristics to converge to a solution with an optimum number of clusters. In this study, ISODATA clustering was performed using the fast implementation of the ISODATA algorithm provided in the SAGA GIS package [48].

*2.5.3. Random Forest Clustering.* As a supervised learning method, random forest classification requires training with a labeled or classified dataset. In this study, we decided to use synthetic training data generated by $k$-means clustering on a random sample of 5000 grid cells of the dataset. The random forest model was trained on synthetic training data, and the dataset was clustered using the trained model. Random forest classification was performed using the randomForest package in *R* [49].

*2.5.4. Number of Clusters.* Although the number of different climate zones is not known a priori, it is a necessary input for clustering methods, such as $k$-means. Some climate clustering studies have adopted a top-down approach to select the number of clusters ($k$) such that five and 13 cluster solutions are derived to match the first two levels of division in KG classification [26]. Various statistical measures, such as the Akaike information criterion, Bayes information criterion, information-theoretical $V$-measure, and Calinski–Harabasz criterion, have been used to determine the optimum number of clusters [42, 50, 51]. In this study, we used the Calinski–Harabasz criterion, which uses the pseudo-$F$ statistic as a measure of cluster cohesiveness [52] because it has been widely used to determine the optimum number of clusters in many applications, including climatological clustering studies [33, 53, 54].

To allow a closer comparison with KG classification, 30-cluster solutions were also developed using both $k$-means and random forest clustering.

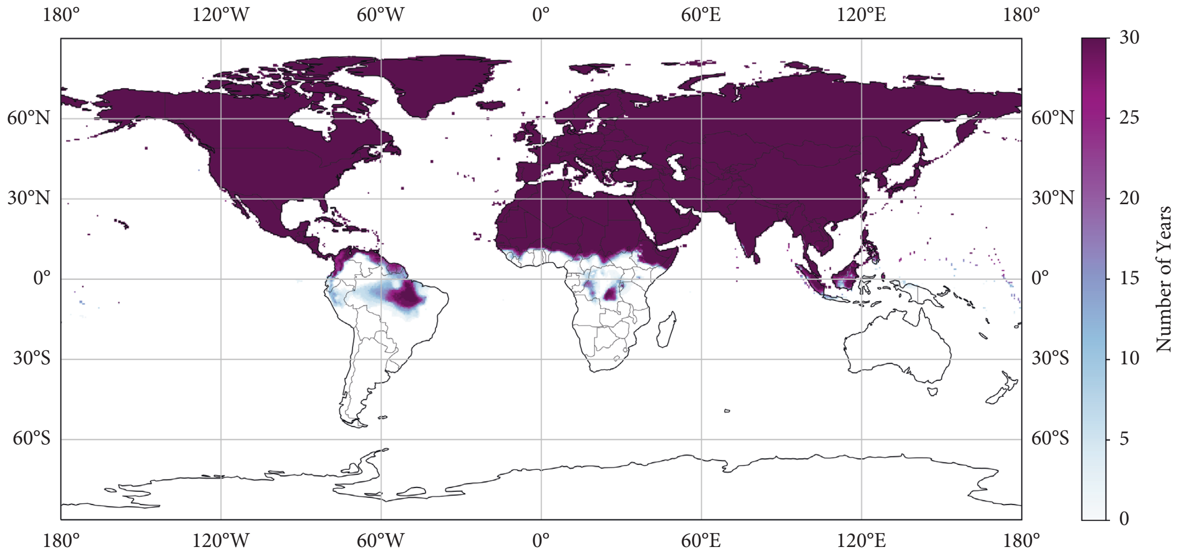


Figure 1: April–September summer regions for 1991–2020.

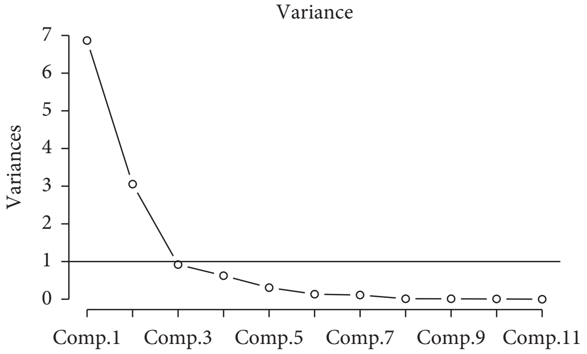


Figure 2: Scree plot showing the eigenvalues for the 11 components of the data variables. Elbow in the plot can be observed at component 3.

Table 1: Loadings of the three principal components.

| Variable | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| MAT | 0.318 | 0.313 | |
| $T_{\text{smin}}$ | 0.311 | 0.322 | |
| $T_{\text{smax}}$ | 0.258 | 0.388 | −0.126 |
| $T_{\text{wmin}}$ | 0.327 | 0.25 | 0.132 |
| $T_{\text{wmax}}$ | 0.316 | 0.316 | |
| MAP | 0.474 | −0.431 | |
| $P_{\text{sdry}}$ | 0.262 | −0.297 | |
| $P_{\text{swet}}$ | 0.303 | −0.17 | 0.13 |
| $P_{\text{wdry}}$ | 0.218 | −0.327 | −0.124 |
| $P_{\text{wwet}}$ | 0.301 | −0.271 | |
| Elevation | | | 0.959 |

*2.5.5. Comparison of Clustering Results.* We used the Jaccard similarity coefficient [55] to investigate the similarity between clusters produced by different methods and KG classes. The Jaccard similarity coefficient is the ratio between the intersection and union of two sets; it has values ranging from zero for non-intersection to one for exact similarity. This index is widely used in the evaluation of similarity in clustering in addition to applications such as image recognition and text analysis [56, 57].

## 3. Results

*3.1. Reproduction of the KG Classification.* A map of the KG classification of the present climate was prepared at a 0.5° resolution (Figure 3). Antarctica was not classified because the CRU TS dataset does not cover the continent. Maps of the five main KG groups and 13 level-2 classes are presented in Figures S2-1 and S2-2 in the Supplementary Materials.

By applying the KG classification scheme at an annual scale for 1901–2020, the annual variation in KG classes was investigated. Figure 4 shows the variability in the five main Köppen climate groups. Maps of variability between individual climate pairs are included in Supplementary Material S3. In the case of the main KG climate groups, two types of areas could be differentiated, as shown in Figure 4. There are narrow and sharp regions that suggest that the corresponding climate groups are well-defined with less ambiguity. There are also wider and fuzzy regions that suggest that the definitions of the corresponding climate groups are ambiguous. The identified regions of high variability agreed well with the findings of Triantafyllou and Tsonis [23]. Although KG is intended to be a classification of long-term climates, its application at an annual scale allows the identification of climates that are prone to be ambiguously characterized.

*3.2. PCA.* The first PC, which represents 57% of the variance, is a combination of all 10 climatic variables, with $P_{\text{year}}$ having largest magnitude. $P_{\text{year}}$ has the largest magnitude in the second PC too, which explains 25% of the variance. Overall, most climatic variables had similar magnitudes in the first two components, suggesting that variance in the original data was shared similarly between the temperature and

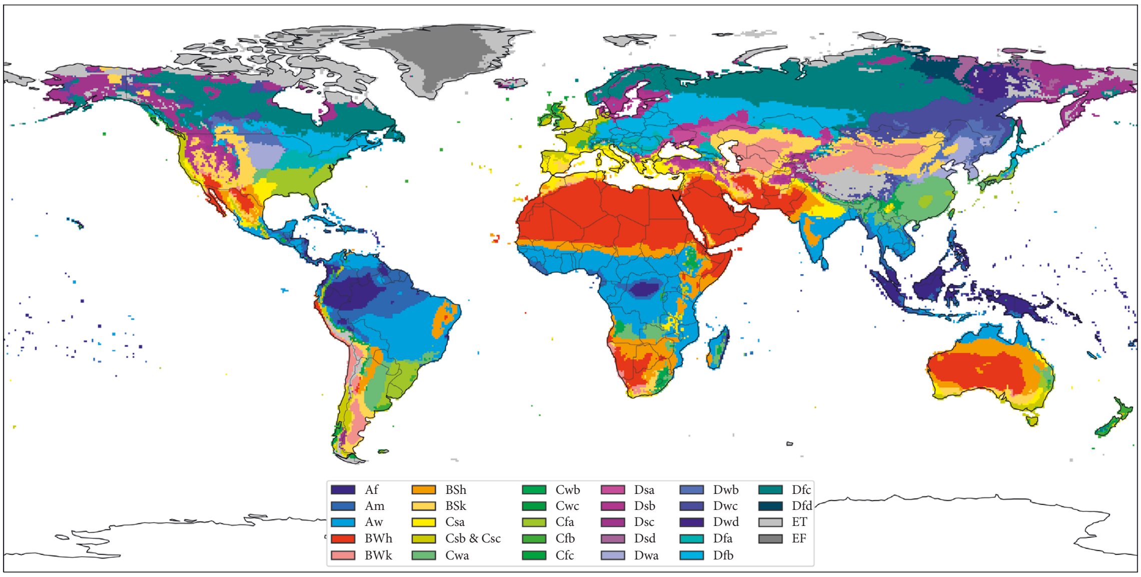


Figure 3: Köppen–Geiger classification of the present world climate. The color scheme has been adopted from Peel et al. [20].

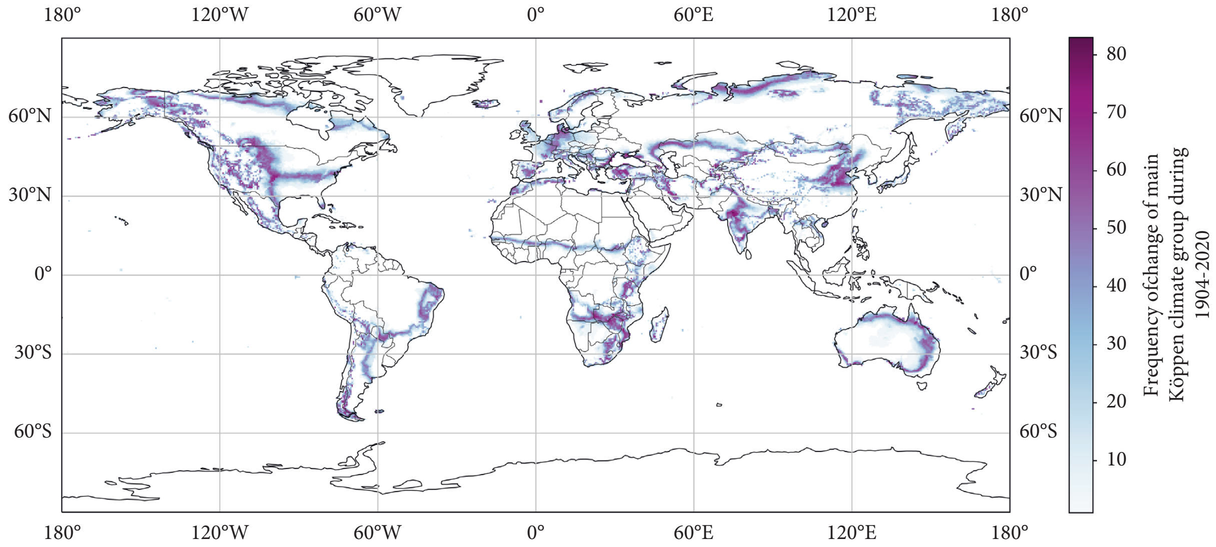

180°
120°W
60°W
0°
60°E
120°E
180°
60°N
30°N
0°
30°S
60°S
80
70
60
50
40
30
20
10
Frequency ofchange of main Köppen climate group during 1904-2020

Figure 4: Frequency of the interannual change of the main Köppen climate class in 1901–2020.

precipitation and the two seasons. The third component, accounting for 7.6% of the variance, was dominated by elevation.

*3.3. Cluster Analysis.* $k$-means clustering was performed on data variables transformed into the three PCs. The pseudo-$F$ statistic was calculated for clustering solutions of up to 35 clusters. An optimum number of clusters ($k = 12$) was selected by detecting the presence of a local peak, followed by a sharp decline in the pseudo-$F$ statistic (Figure 5). Based on this, a 12-cluster solution (KM12) was prepared using $k$-means clustering, as shown in Figure 6. Then, a 30-cluster solution (KM30) was prepared (Figure 7). ISODATA clustering resulted in a 16-cluster solution (Figure 8) named ISO16. Random forest clustering was used to develop a 30-cluster solution (Figure 9) named RF30. All cluster maps were visualized using the same color scale applied in the order of the mean annual temperature of each cluster. Cluster identification numbers were assigned in the same order.

*3.4. Jaccard Coefficient.* The Jaccard coefficient between the KG classification and each data-driven classification was calculated. The Jaccard coefficient values are listed in Supplementary Material S4.

## 4. Discussion

*4.1. Discussion of Results.* Visual observation revealed several noticeable similarities and differences between the four data-driven classifications and the KG classification. All four clustering schemes created larger clusters with similar

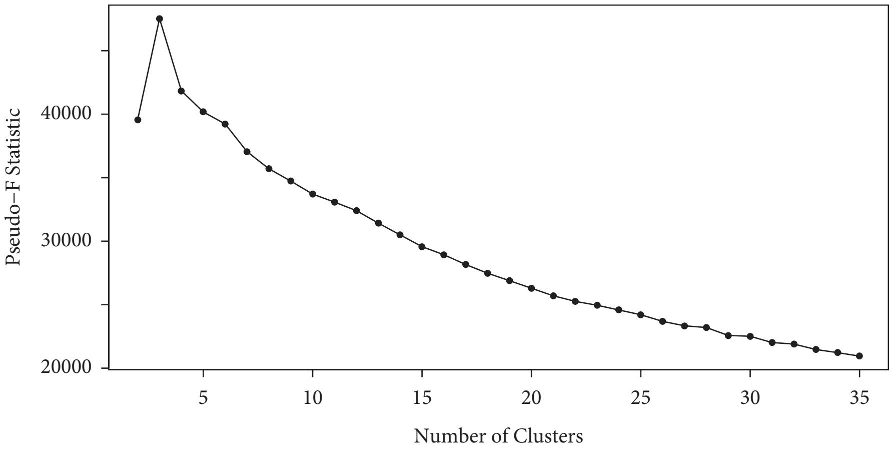


Figure 5: Pseudo-$F$ statistic for clustering with different number of clusters ($k$).

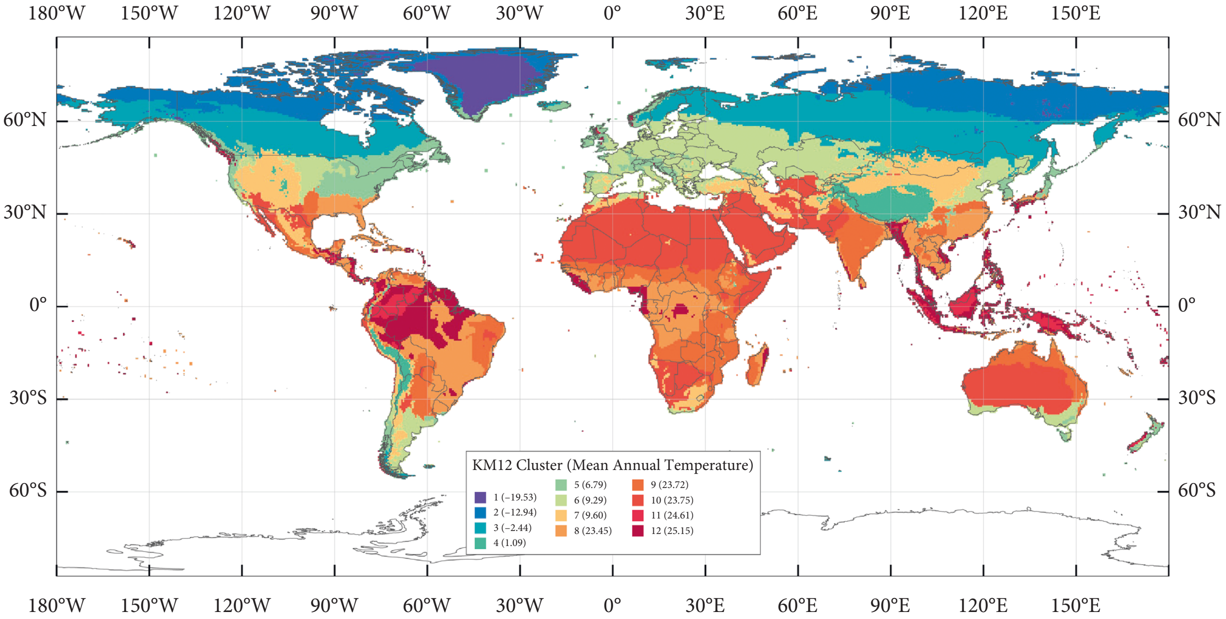


Figure 6: 12-cluster solution (KM12) of $k$-means clustering.

boundaries in some areas of the world, such as in Northern and Central Africa, Eastern Americas, Northern and Central Eurasia, and Australia. In these regions, data-driven clusters have visual similarities to the KG climate classes. Other regions, such as western Americas, East Africa, and East and Southeast Asia, generally have smaller and more fragmented clusters. For a fair assessment, KM12 and ISO16 can be compared with the 13 level-2 KG climate classes, and KM30 and RF30 can be compared with the full 30-class KG classification.

*4.2. Cluster Similarities.* In all four clustering results, there are clusters similar to some KG climates, particularly in groups *A*, *B*, and *E*, which are defined primarily by more extreme temperatures. Similarities can be identified by inspecting the Jaccard coefficients (Supplementary Material S4). Further, the proportions of coverage between different classification schemes are visualized in Figure 10.

Between the $k$-means 12-cluster solution and the 13 level-2 KG climate classes, the highest similarity is present in the EF KG climate to KM12 cluster 1 and BWh KG climate to KM12 cluster 10, with Jaccard similarity coefficient values of 0.81 and 0.67, respectively. Figure 11 shows how some of the regional boundaries are in close proximity, signaling that KG is able to identify some of the natural clusters in climate data. The dissimilarity between the BW KG class and KM12 cluster 10 is mainly due to the cold desert regions that correspond to the BWk class and are not included in KM12 cluster 10.

Figure 12 plots the mean values of the two main PCs for the 13 level-2 KG climates and the KM12 clusters. Some centers are close together indicating similar regions in the two classification systems. At the same time, cluster analysis has recognized some unique climatic regions that were not seen in the KG classification results, as indicated by the presence of several isolated cluster centers.

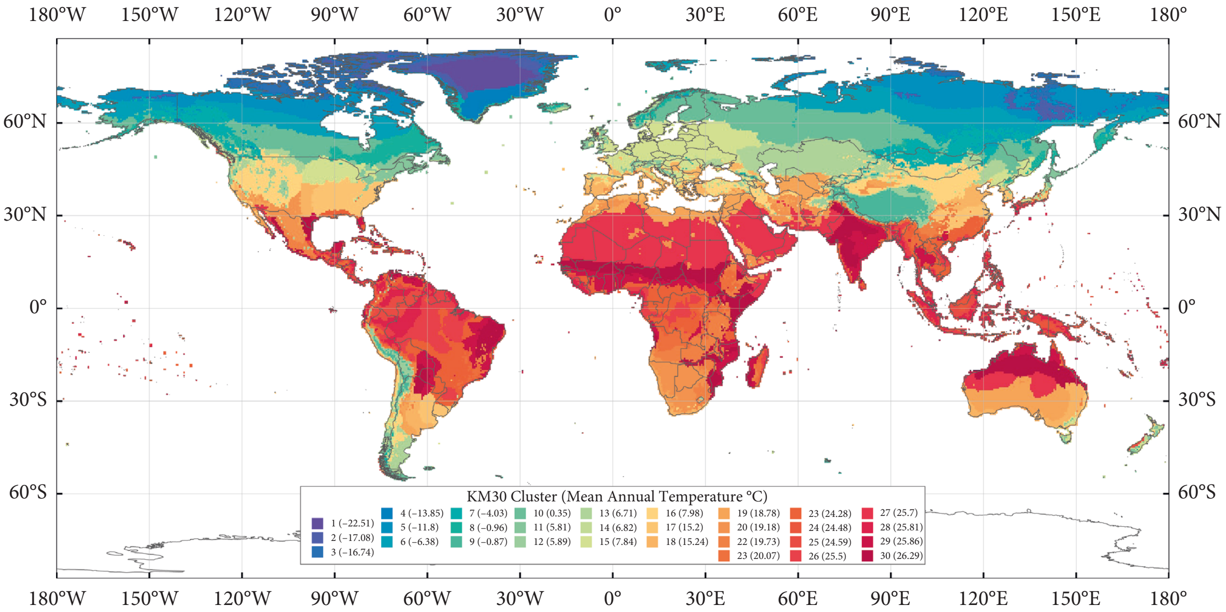


Figure 7: 30-cluster solution (KM30) of *k*-means clustering.

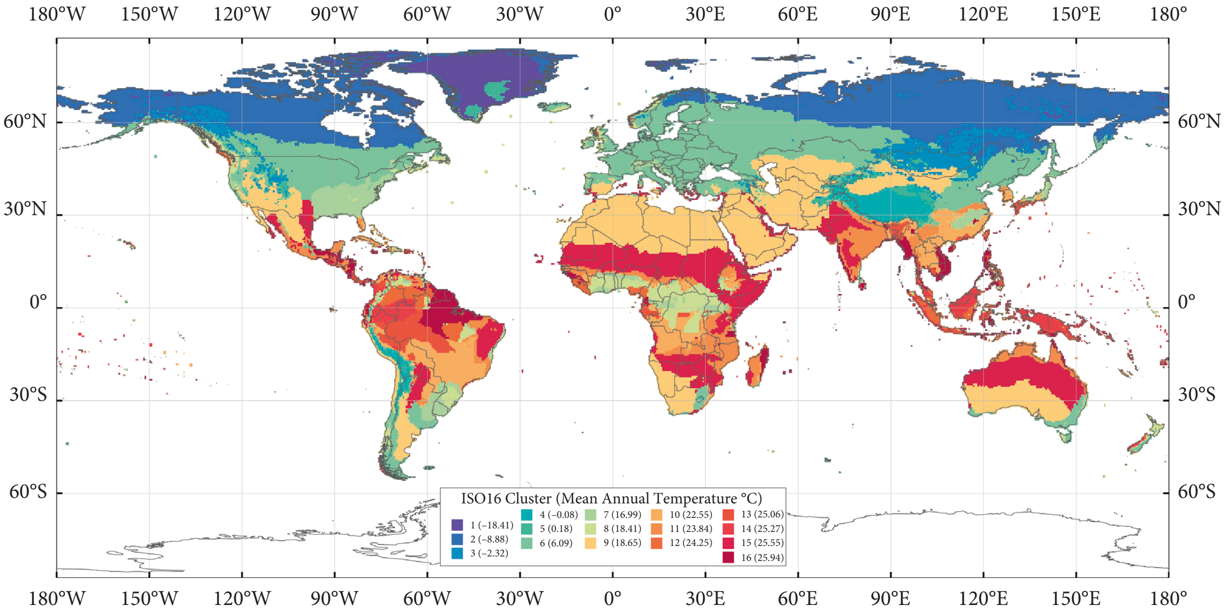


Figure 8: 16 ISODATA clusters (ISO16).

Cluster 11 of the KM12 stands out in Figure 12. In the KG, that area is shared between Af, Am, and Cf classes with Jaccard coefficients of 0.446, 0.035, and 0.13, respectively.

The relative contribution of the parameters for the clustering result can be studied based on the standard deviation. In distance-based *k*-means clustering and ISODATA clustering methods, the objective of the algorithm is to minimize the distance between cluster members and cluster mean, given by the within-cluster sum of squares (WCSS), which is equivalent to variance. Therefore, by calculating the standard deviation of each variable of the cluster members (Figure 13 for KM12), the contribution of the variables to the minimization of the objective function of WCSS can be compared. A variable that has high similarity, i.e., low standard deviation, contributes more to the differentiation of the clusters.

The contribution of the precipitation variables is significantly high in some clusters like 2, 3, and 10 but significantly low in clusters 11 and 12. The influence of temperature variables appears to be similar for all clusters. This distinction can be explained by the high dynamic range of original precipitation variables. Contribution of elevation is comparatively low across all clusters.

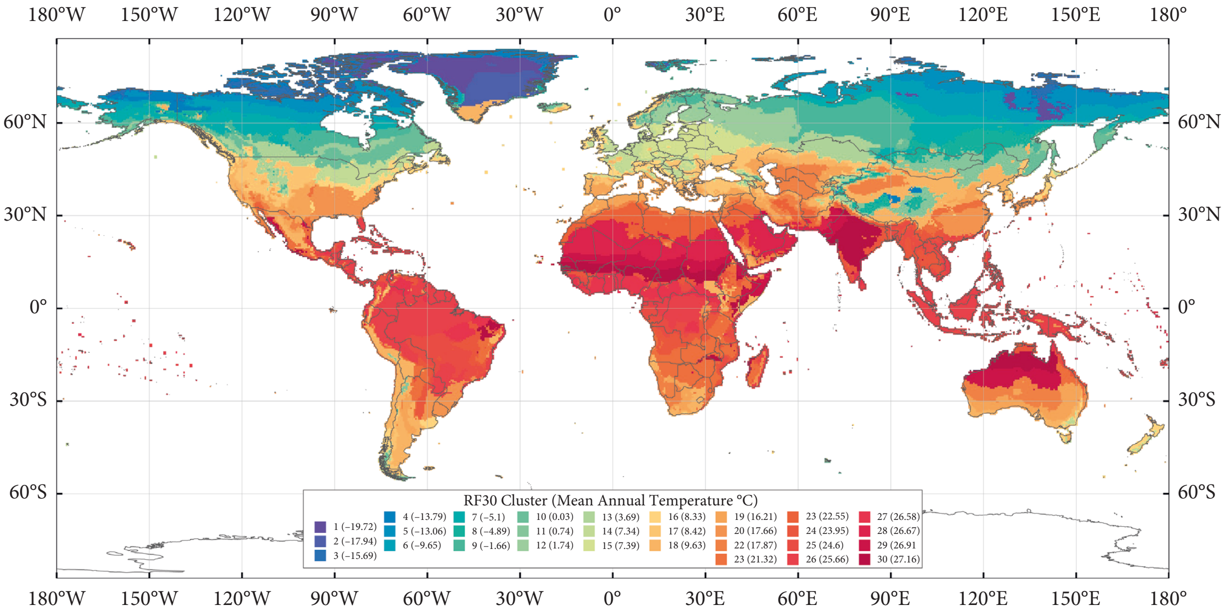


Figure 9: 30-cluster solution (RF30) of random forest clustering.

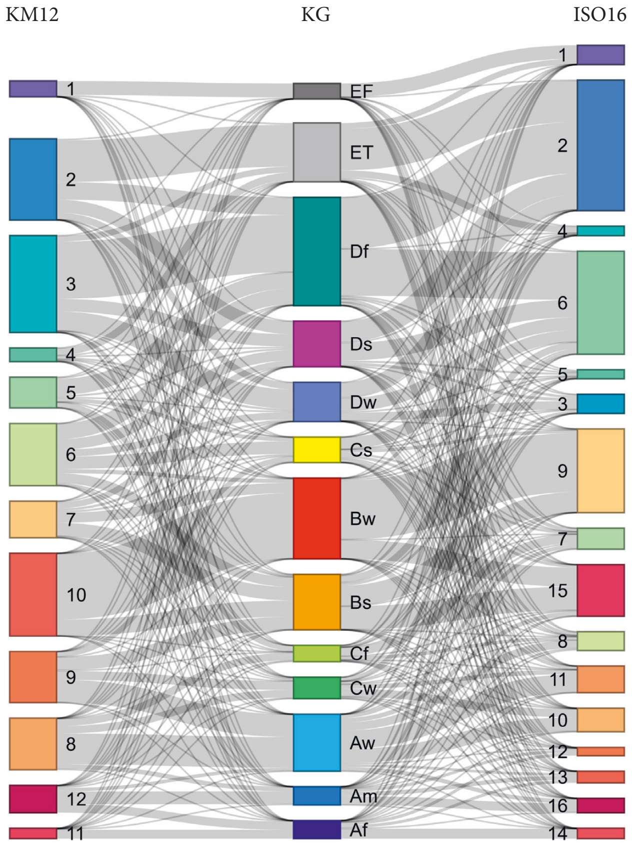


Figure 10: Proportions of coverage of climate regions across $k$-means 12 clusters, Köppen–Geiger level 2, and ISODATA 16-cluster classification schemes.

*4.3. Climatic Regions Discovered by Cluster Analysis.* There are instances wherein cluster analysis has subdivided the Köppen classes. The Af KG climate for tropical rainforests has no level-3 subdivision in KG. In the KM12 classification, the region corresponding to the Af climate was shared mainly between two clusters, numbers 11 and 12. As revealed by the climographs (Figure 14) of the regions, $k$-means clustering distinguished different precipitation levels, although the temperature variation was similar in all three regions. KM12 cluster 11 has distinctly higher precipitation than KM12 cluster 12. Noting that, in KG, the Af class is defined by minimum monthly precipitation >60 mm, we can investigate the value of the same statistic for the KM12 clusters. The mean of the minimum monthly precipitation is 132.84 mm in KM12 cluster 11 and 48.11 mm in KM12 cluster 12. The mean annual precipitation is 3368 mm in cluster 11 and 2360 mm in cluster 12. These statistics can be interpreted as follows. When identifying the wettest climate, KM12 establishes a lower threshold for precipitation that is more than twice the value in KG. While the Af class contains all three major rainforest regions in Central Africa, South America, and Southeast Asia, cluster 11 is not present in the African continent.

The 30-cluster solutions from $k$-means and random forest clustering provided further insights into refining the KG climate classes. The warm summer continental climate of Dfb, which is present in Eurasia and North America, has been placed in separate clusters in the random forest classification (Figure 15). Climographs revealed that the North American part of the Dfb class receives distinctly higher precipitation, suggesting that the Dfb class can be subdivided further.

Cluster analysis has detected isolated geographical areas with unique climates that were hidden in the KG classification results. For instance, both 30-cluster solutions KM30 and RF30 distinguish the Tibesti mountain region in central Sahara from the surrounding Sahara Desert. This new cluster is present in other arid areas of the world, such as the mountainous regions of West Sahara, southern Africa, Middle East, and southwestern North America. The climate

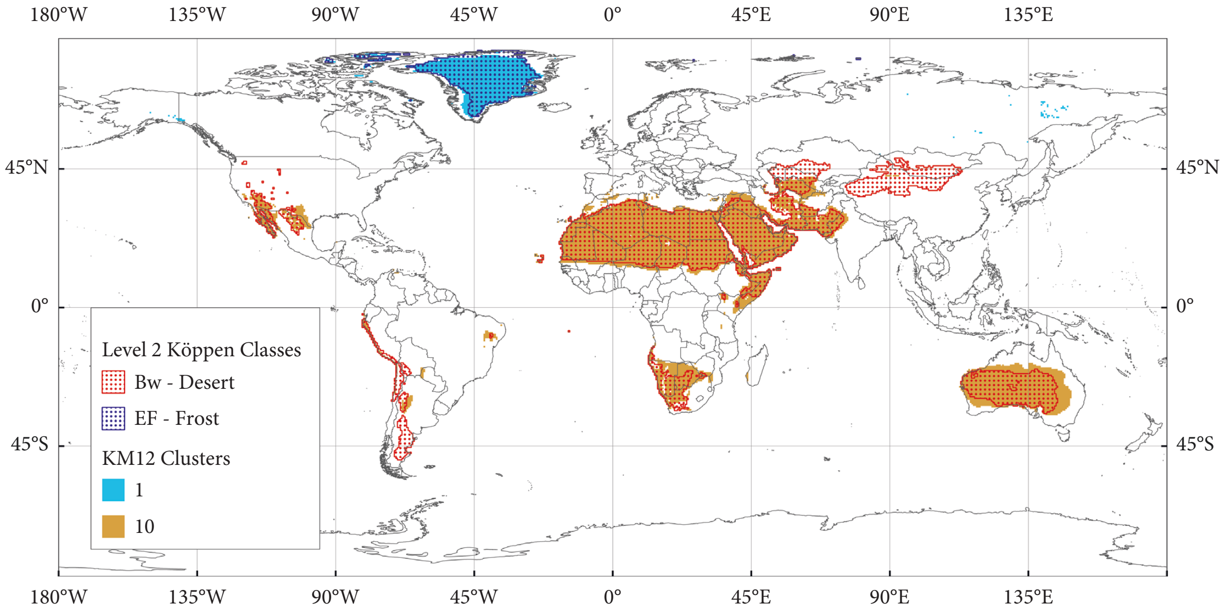


Figure 11: Köppen classes and *k*-means clusters with the highest similarity. There are clusters in *k*-means 12-cluster results that closely resemble the EF polar climate and the Bwh desert climate. Cold desert regions of the Bwh class have been excluded from the corresponding *k*-means cluster.

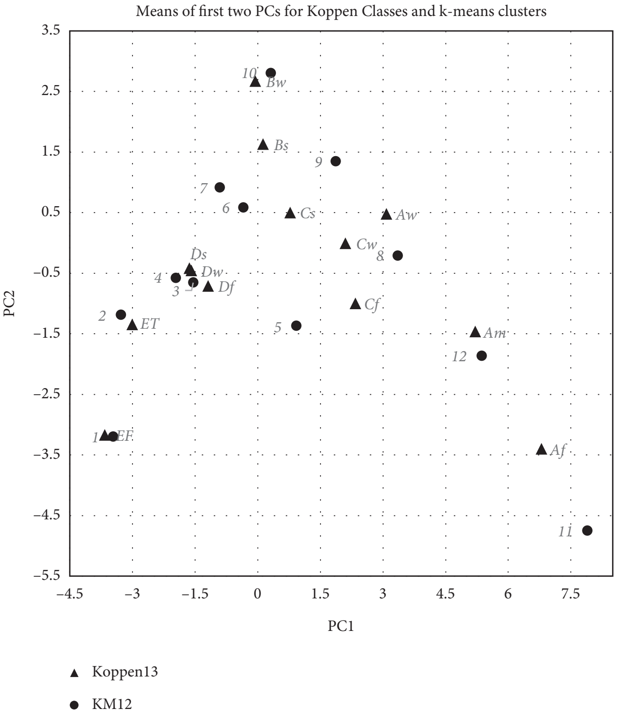
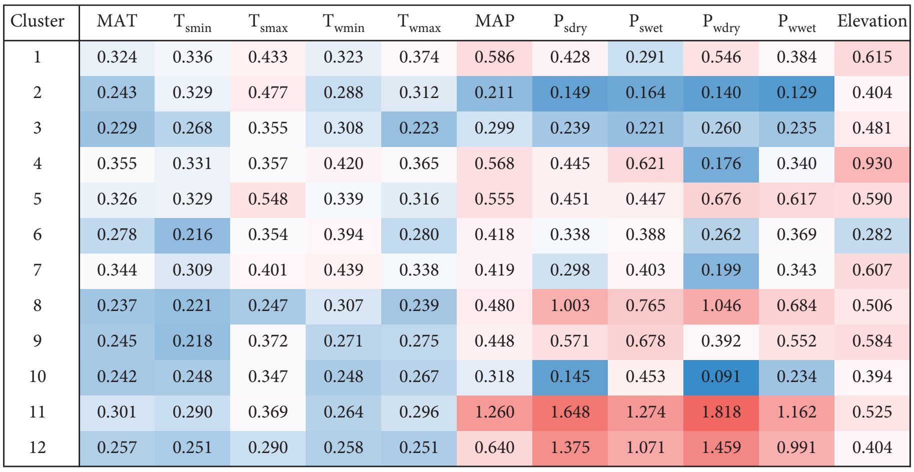


Figure 12: Means of the two PCs for the 13 level-2 KG climates (Köppen13) and 12 *k*-means clusters (KM12). Centers of some clusters (5, 7, and 11) are distant from any KG class means.

in the Tibesti mountains is different from that in the rest of the Sahara, as it receives more rainfall. Figure 16 shows that cluster 20 of the KM30 classification, which includes the above areas, receives substantially higher rainfall, in contrast to cluster 27, which is more similar to the BWh—hot desert KG climate.

*4.4. Clustering with Random Forest.* Some clusters created by the distance-based *k*-means and ISODATA clustering methods appear to be primarily sensitive to either temperature or precipitation, rather than both. For instance, Clusters 11 and 12 of KM12 (which are the two warmest clusters), cluster 25 of KM30 (which is the sixth warmest cluster), and cluster 14 of ISO16 (which is the third warmest) have parts in colder regions, such as Scotland, Norway, New Zealand, and Chile, corresponding to locations of the boreal rainforests that receive very high precipitation. These regions are clustered together with tropical rainforest regions, even though the temperature regimes are considerably different. However, forest-based clustering did not show a similar outcome. In RF30, the same higher precipitation regions at higher latitudes were clustered together with other colder regions. This highlights some advantages of machine learning models, such as random forest and other decision tree-based models, over distance-based clustering methods. In particular, compared with distance-based methods, tree-based methods are generally more robust against outliers [58].

*4.5. Suggestions for Dissemination of Data-Driven Classification Results.* Although data-driven classification offers an objective classification scheme with superfluous details, wider adoption of such schemes may be discouraged by certain common drawbacks. The climate zones produced by data-driven methods need to be retrospectively characterized. Not only may it be challenging to uniquely characterize all individual clusters, but a single definition may also not be satisfactory for users in different disciplines. Another problem is that the clustering results can be inconsistent. For instance, because the *k*-means algorithm converges to a local



| Cluster | MAT | $T_{smin}$ | $T_{smax}$ | $T_{wmin}$ | $T_{wmax}$ | MAP | $P_{sdry}$ | $P_{swet}$ | $P_{wdry}$ | $P_{wwet}$ | Elevation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.324 | 0.336 | 0.433 | 0.323 | 0.374 | 0.586 | 0.428 | 0.291 | 0.546 | 0.384 | 0.615 |
| 2 | 0.243 | 0.329 | 0.477 | 0.288 | 0.312 | 0.211 | 0.149 | 0.164 | 0.140 | 0.129 | 0.404 |
| 3 | 0.229 | 0.268 | 0.355 | 0.308 | 0.223 | 0.299 | 0.239 | 0.221 | 0.260 | 0.235 | 0.481 |
| 4 | 0.355 | 0.331 | 0.357 | 0.420 | 0.365 | 0.568 | 0.445 | 0.621 | 0.176 | 0.340 | 0.930 |
| 5 | 0.326 | 0.329 | 0.548 | 0.339 | 0.316 | 0.555 | 0.451 | 0.447 | 0.676 | 0.617 | 0.590 |
| 6 | 0.278 | 0.216 | 0.354 | 0.394 | 0.280 | 0.418 | 0.338 | 0.388 | 0.262 | 0.369 | 0.282 |
| 7 | 0.344 | 0.309 | 0.401 | 0.439 | 0.338 | 0.419 | 0.298 | 0.403 | 0.199 | 0.343 | 0.607 |
| 8 | 0.237 | 0.221 | 0.247 | 0.307 | 0.239 | 0.480 | 1.003 | 0.765 | 1.046 | 0.684 | 0.506 |
| 9 | 0.245 | 0.218 | 0.372 | 0.271 | 0.275 | 0.448 | 0.571 | 0.678 | 0.392 | 0.552 | 0.584 |
| 10 | 0.242 | 0.248 | 0.347 | 0.248 | 0.267 | 0.318 | 0.145 | 0.453 | 0.091 | 0.234 | 0.394 |
| 11 | 0.301 | 0.290 | 0.369 | 0.264 | 0.296 | 1.260 | 1.648 | 1.274 | 1.818 | 1.162 | 0.525 |
| 12 | 0.257 | 0.251 | 0.290 | 0.258 | 0.251 | 0.640 | 1.375 | 1.071 | 1.459 | 0.991 | 0.404 |

Figure 13: Standard deviation of the normalized variables in each cluster of *k*-means 12-cluster solution.

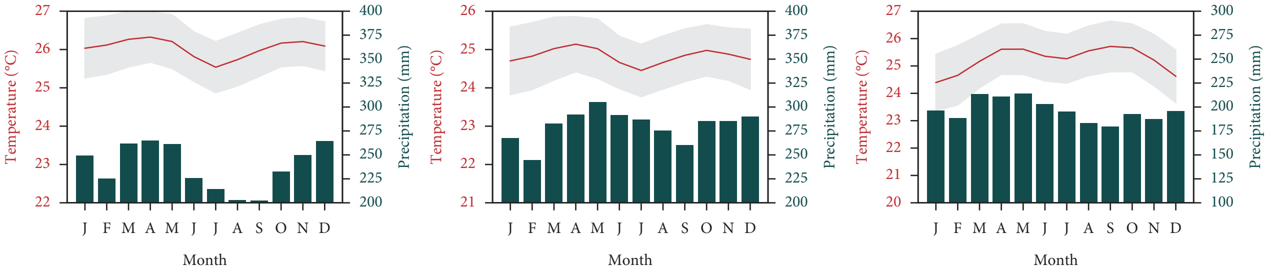


Figure 14: Climographs of KG class Aw and the two KM12 clusters that share the corresponding area (left: Aw tropical rainforest KG climate; center: KM12 cluster 11; right: KM12 cluster 12). Temperature variations are similar in all three, whereas precipitation is different in the two *k*-means clusters.

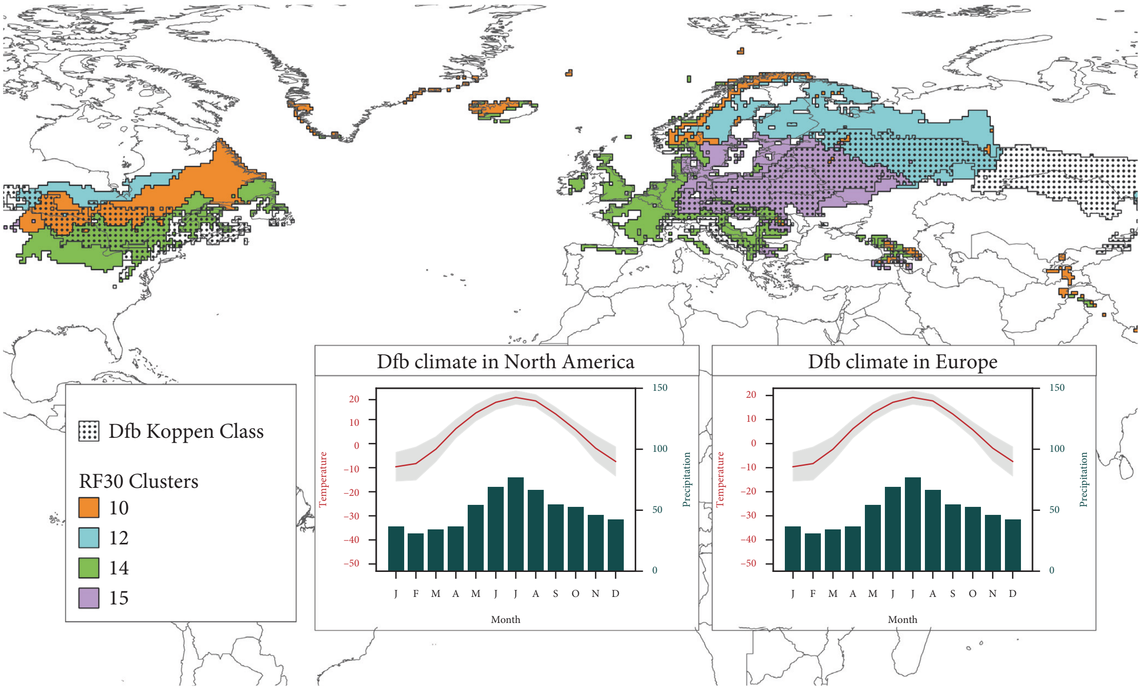


Figure 15: Dfb warm summer continental KG class and the overlapping RF30 clusters. Inset climographs are for the two parts of Dfb classes in Eurasia and North America.

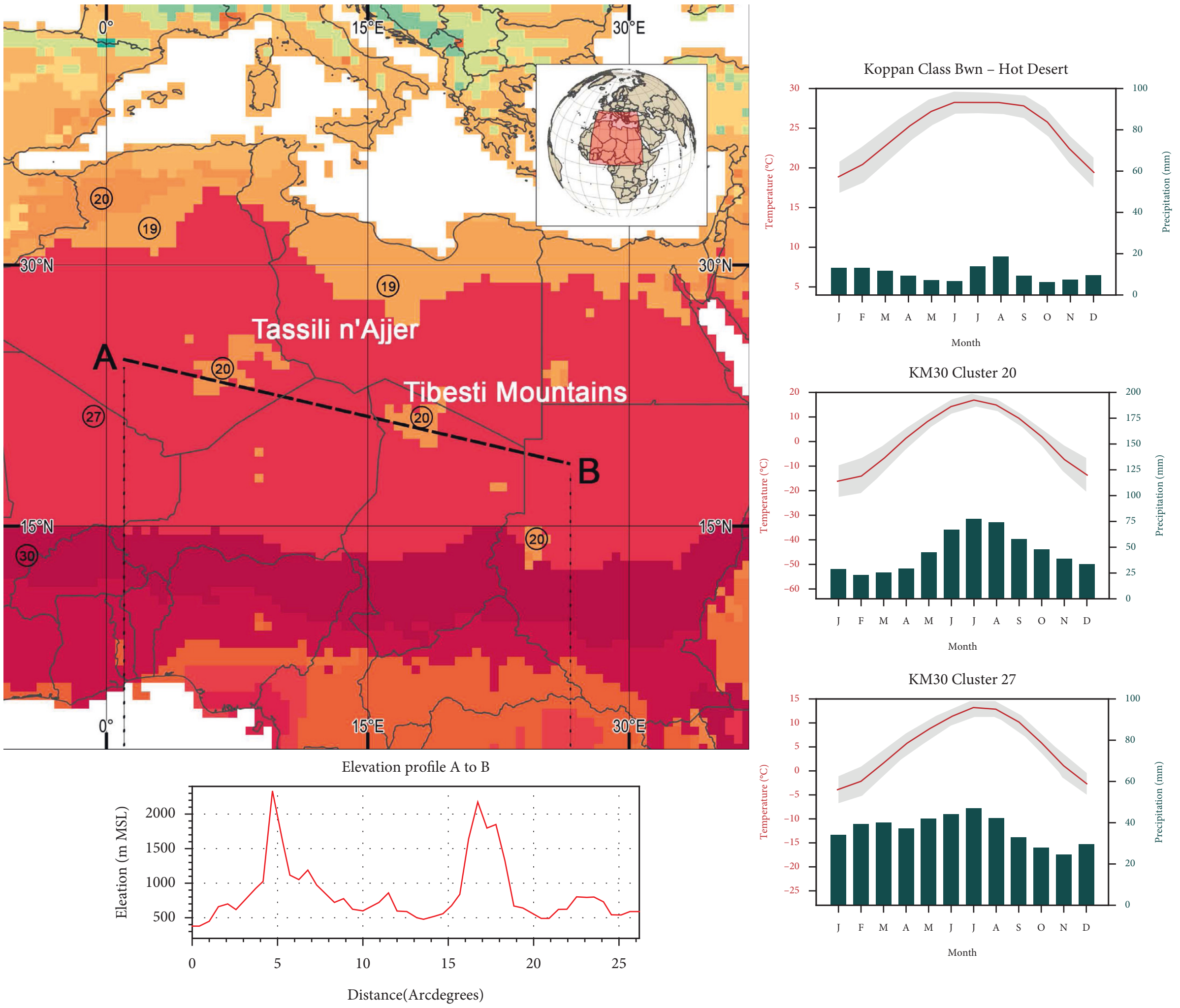


Figure 16: Left: Tibesti and Tassili n'Ajjer mountain regions in the Sahara. Right: Climographs of Bwh—hot desert Köppen climate and clusters 20 and 27 of KM30 classification. Bottom: elevation profile from *A* to *B* showing higher elevation regions within the Sahara.

minimum, it can produce different clusters in each trial with different initial conditions [59]. Although methods to search for optimum initial conditions [60, 61] and reproduce a selected clustering outcome are available, the above attributes may make data-driven classifications less appealing than the intuitive and unique results offered by rule-based classifications.

Maps of climate zones produced by data-driven methods are highly fragmented with complex edges in some regions. For the ease of communication, maps of data-driven classifications can be post-processed to reduce noise and sharpen edges. Denoising is a process employed in signal and image processing to remove noise from analog and digital signals and images. There are many different image denoising techniques [62]. For application in climate classification maps, a suitable technique can be chosen considering the requirements that the pixel values in the resulting filtered image must be limited to the values (cluster identification) in the original image and that edges must be preserved and sharpened. Rank filtering satisfies both these requirements. It replaces a pixel value with a specified ranking value from the sorted values of the neighborhood. Often, the rank is specified as the median, and this process is called a median filter [63]. Supplementary Material S5 shows a map of the KM30 clusters post-processed using a median filter. The climate zones in the denoised map are less fragmented, have sharper edges, and are generally more discernible. Therefore, such post-processing techniques can be used to produce maps that are better suited for the communication of the results of data-driven classification.

## 5. Concluding Remarks

While the rule-based KG classification system has been well established as the foremost climate classification system, data-driven classifications offer an alternative with the promise of being more objective. Our study was devised to explore naturally emerging clusters in climate data and compare the identified climatic regions with those obtained with the KG classification system. Global climatic regions were objectively delineated by conducting cluster analyses on a data matrix comprising 10 climatic variables and elevation.

In the climatic regions identified by cluster analyses, strong similarities to KG climate classes were observed in regions with extreme temperatures. All clustering methods delineated prominent climate regions that were similar to the KG climate classes in groups *A* (tropical), *B* (arid), and *C* (polar). Higher Jaccard coefficient values were also reported among the above groups, confirming that the best consensus between data-driven classifications and KG exists in these climates. In the temperate (*C*) and cold (*D*) KG climate groups, agreement with data-driven classifications was limited.

Further refinements to the KG climate classes were suggested based on the results of data-driven clustering. Instances wherein KG climate classes could be subdivided into distinct climates were identified, such as the subdivision of the Af-tropical rainforest climate and the Dfb warm summer continental climate. Unique climatic regions that were obscured in KG were also identified, such as mountainous regions within the Sahara.

In summary, our clustering results show that even though it is a rule-based classification system, KG approximates some of the natural clusters in terms of the global climate. Simultaneously, it obscures regions that can be differentiated in data-driven classifications. With no definitive measure of the performance of climate classification systems, it is impossible to conclude that one system is better than the other. Clustering-based climate classifications may be less appealing as a stand-alone system because of the lack of formal definitions, whereas KG has established a wide appeal due to its familiar definitions. However, definitions for data-driven clusters can be formulated based on climatology and geography, as demonstrated in selected cases.

In addition, we conclude that data-driven classifications are best used to complement and refine the structure and definition provided by the rule-based KG classification system. To that end, we demonstrated how Köppen classes can be refined using data-driven insights. In addition, post-processing methods, such as demonstrated median filtering, may be suitable for developing climate zone maps that are suitable for interpretation and communication.

The climate data selected in this study were limited to monthly means of the temperatures and precipitation. There is an opportunity to enrich data-driven classifications by including more variables that are descriptive or predictive of the climate. Further studies should focus on investigating additional variables that could produce more insightful clustering outputs.

## Data Availability

Results of climate regionalization and updated Köppen–Geiger classification can be accessed at https://doi.org/10.5281/zenodo.6360920.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This research was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS), in collaboration with Japan Science and Technology Agency (JST) (JPMJSA1502) and Japan International Cooperation Agency (JICA), and the grant KAKENHI (16H06291 and 21H05002) of the Japan Society for the Promotion of Science.

## Supplementary Materials

S1: Köppen classification system and the defining criteria. S2: maps of main and level-2 Köppen climates. S3: variability of Köppen climate classes. S4: Jaccard coefficients. S5: KM30 classification post-processed with median filter. (*Supplementary Materials*)

## References

[1] M. J. Metzger, D. J. Brus, R. G. H. Bunce et al., “Environmental stratifications as the basis for national, European and global ecological monitoring,” *Ecological Indicators*, vol. 33, pp. 26–35, 2013.

[2] M. Ortega, M. J. Metzger, R. G. H. Bunce et al., “The potential for integration of environmental data from regional stratifications into a European monitoring framework,” *Journal of Environmental Planning and Management*, vol. 55, no. 1, pp. 39–57, 2012.

[3] H. Frischkorn, J. C. Araújo, and M. M. F. Santiago, “Water resources of ceará and piauí,” in *Global Change and Regional Impacts*, T. Gaiser, M. Krol, H. Frischkorn, and J. C. de Araújo, Eds., Springer, Berlin, Germany, 2003.

[4] W. J. M. Knoben, R. A. Woods, and J. E. Freer, “A quantitative hydrological climate classification evaluated with independent streamflow data,” *Water Resources Research*, vol. 54, no. 7, pp. 5088–5109, 2018.

[5] S. D. Maddux, T. R. Yokley, B. M. Svoma, and R. G. Franciscus, “Absolute humidity and the human nose: a reanalysis of climate zones and their influence on nasal form and function,” *American Journal of Physical Anthropology*, vol. 161, no. 2, pp. 309–320, 2016.

[6] A. Araya, S. D. Keesstra, and L. Stroosnijder, “A new agro-climatic classification for crop suitability zoning in northern semi-arid Ethiopia,” *Agricultural and Forest Meteorology*, vol. 150, no. 7-8, pp. 1057–1064, 2010.

[7] S. J. Bacon, A. Aebi, P. Calanca, and S. Bacher, “Quarantine arthropod invasions in Europe: the role of climate, hosts and propagule pressure,” *Diversity and Distributions*, vol. 20, no. 1, pp. 84–94, 2014.

[8] J. Tonietto and A. Carbonneau, “A multicriteria climatic classification system for grape-growing regions worldwide,” *Agricultural and Forest Meteorology*, vol. 124, no. 1-2, pp. 81–97, 2004.

[9] M. B. Mathur, R. B. Patel, M. Gould et al., “Seasonal patterns in human A (H5N1) virus infection: analysis of global cases,” *PLoS One*, vol. 9, no. 9, Article ID e106171, 2014.

[10] F. Méndez-Arriaga, “The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1,” *Science of the Total Environment*, vol. 735, Article ID 139560, 2020.

[11] J. J. Feddema, "A revised thornthwaite-type global climate classification," *Physical Geography*, vol. 26, no. 6, pp. 442–466, 2005.

[12] N. B. Guttman and R. G. Quayle, "A historical perspective of U.S. Climate divisions," *Bulletin of the American Meteorological Society*, vol. 77, no. 2, pp. 293–303, 1996.

[13] L. R. Holdridge, *Life Zone Ecology*, Tropical Science Center, San Jose, CA, USA, 1967.

[14] W. Köppen, "Das geogr syst klimate," *Handbuch der Klimatologie*, Borntraeger, Berlin, Germany, 1936.

[15] M. Sanderson, "The classification of climates from pythagoras to Koeppen," *Bulletin of the American Meteorological Society*, vol. 80, no. 4, pp. 669–673, 1999.

[16] R. Geiger and W. Pohl, "Eine neue Wandkarte der Klimagebiete der Erde nach W. köppens klassifikation (a new wall map of the climatic regions of the world according to W. Köppen's classification)," *Erdkunde*, vol. 8, pp. 58–61, 1954.

[17] G. T. Trewartha, *An Introduction to Climate*, McGraw Hill, New York, NY, USA, 1968.

[18] J. Grieser, R. Gommes, S. Cofield, and M. Bernardi, *FAO Climate Maps*, FAO, Rome, Italy, 2006.

[19] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World Map of the Köppen-Geiger climate classification updated," *Meteorologische Zeitschrift*, vol. 15, no. 3, pp. 259–263, 2006.

[20] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Koppen-Geiger climate classification," *Hydrology and Earth System Sciences*, vol. 12, 2007.

[21] F. Rubel and M. Kottek, "Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification," *Meteorologische Zeitschrift*, vol. 19, no. 2, pp. 135–141, 2010.

[22] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and future Köppen-Geiger climate classification maps at 1 km resolution," *Scientific Data*, vol. 5, no. 1, Article ID 180214, 2018.

[23] G. N. Triantafyllou and A. A. Tsonis, "Assessing the ability of the Köppen system to delineate the general world pattern of climates," *Geophysical Research Letters*, vol. 21, no. 25, pp. 2809–2812, 1994.

[24] F. Rubel and M. Kottek, "Comments on: "the thermal zones of the earth" by Wladimir köppen (1884)," *Meteorologische Zeitschrift*, vol. 20, no. 3, pp. 361–365, 2011.

[25] M. J. Metzger, R. G. H. Bunce, R. H. G. Jongman, R. Sayre, A. Trabucco, and R. Zomer, "A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring: high-resolution bioclimate map of the world," *Global Ecology and Biogeography*, vol. 22, no. 5, pp. 630–638, 2013.

[26] P. Netzel and T. Stepinski, "On using a clustering approach for global climate classification," *Journal of Climate*, vol. 29, no. 9, pp. 3387–3401, 2016.

[27] M. T. Tercek, S. T. Gray, and C. M. Nicholson, "Climate zone delineation: evaluating approaches for use in natural resource management," *Environmental Management*, vol. 49, no. 5, pp. 1076–1091, 2012.

[28] J. Zscheischler, M. D. Mahecha, and S. Harmeling, "Climate classifications: the value of unsupervised clustering," *Procedia Computer Science*, vol. 9, pp. 897–906, 2012.

[29] A. T. DeGaetano, "Spatial grouping of United States climate stations using a hybrid clustering approach," *International Journal of Climatology*, vol. 21, no. 7, pp. 791–807, 2001.

[30] J. S. Russell and A. W. Moore, "Classification of climate by pattern analysis with Australasian and southern African data as an example," *Agricultural Meteorology*, vol. 16, no. 1, pp. 45–70, 1976.

[31] Y. Unal, T. Kindap, and M. Karaca, "Redefining the climate zones of Turkey using cluster analysis," *International Journal of Climatology*, vol. 23, no. 9, pp. 1045–1055, 2003.

[32] R. G. Fovell and M.-Y. C. Fovell, "Climate zones of the conterminous United States defined using cluster analysis," *Journal of Climate*, vol. 6, no. 11, pp. 2103–2135, 1993.

[33] M. L. Marston and A. W. Ellis, "Delineating precipitation regions of the contiguous United States from cluster analyzed gridded data," *Annals of the Association of American Geographers*, vol. 111, pp. 1–19, 2020.

[34] S. Park, H. Park, J. Im et al., "Delineation of high resolution climate regions over the Korean Peninsula using machine learning approaches," *PLoS One*, vol. 14, no. 10, Article ID e0223362, 2019.

[35] F. M. Hoffman, W. W. Hargrove, D. J. Erickson, and R. J. Oglesby, "Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models," *Earth Interactions*, vol. 9, no. 10, pp. 1–27, 2005.

[36] J. Kumar, R. T. Mills, F. M. Hoffman, and W. W. Hargrove, "Parallel *k*-means clustering for quantitative ecoregion delineation using large data sets," *Procedia Computer Science*, vol. 4, pp. 1602–1611, 2011.

[37] I. Harris, T. J. Osborn, P. Jones, and D. Lister, "Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset," *Scientific Data*, vol. 7, no. 1, p. 109, 2020.

[38] J. J. Danielson and D. B. Gesch, *Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010)*, U.S. Geo-Logical Survey, Sioux Falls, SD, USA, 2011.

[39] D. J. Kriticos, B. L. Webber, A. Leriche et al., "CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling," *Methods in Ecology and Evolution*, vol. 3, no. 1, pp. 53–64, 2012.

[40] R. G. Fovell, "Consensus clustering of U.S. temperature and precipitation data," *Journal of Climate*, vol. 10, no. 6, pp. 1405–1427, 1997.

[41] J. Gómez-Zotano, J. Alcántara-Manzanares, E. Martínez-Ibarra, and J. A. Olmedo-Cobo, "Applying the technique of image classification to climate science: the case of Andalusia (Spain)," *Geographical Research*, vol. 54, no. 4, pp. 461–470, 2016.

[42] K. Kozjek, M. Dolinar, and G. Skok, "Objective climate classification of Slovenia," *International Journal of Climatology*, vol. 37, pp. 848–860, 2017.

[43] D. A. Jackson, "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches," *Ecology*, vol. 74, no. 8, pp. 2204–2214, 1993.

[44] G. Hamerly and J. Drake, "Accelerating lloyd's algorithm for *k*-means clustering," in *Partitional Clustering Algorithms*, M. E. Celebi, Ed., Springer International Publishing, Berlin, Germany, 2015.

[45] J. Macqueen in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*Some methods for classification and analysis of multivariate observations, Berkeley, CA, USA, July 1967.

[46] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *Cluster: Cluster Analysis Basics and Extensions*, 2021.

[47] A. M. Bagirov and E. Mohebi, "Nonsmooth optimization based algorithms in cluster analysis," in *Partitional Clustering Algorithms*, M. E. Celebi, Ed., Springer International Publishing, Berlin, Germany, 2015.

[48] N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, and J. Le Moigne, “A fast implementation of the isodata clustering algorithm,” *International Journal of Computational Geometry and Applications*, vol. 17, no. 1, pp. 71–103, 2007.

[49] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[50] J. Nowosad and T. F. Stepinski, “Spatial association between regionalizations using the information-theoretical *V*-measure,” *International Journal of Geographical Information Science*, vol. 32, no. 12, pp. 2386–2401, 2018.

[51] D. Sathiaraj, X. Huang, and J. Chen, “Predicting climate types for the continental United States using unsupervised clustering techniques,” *Environmetrics*, vol. 30, no. 4, 2019.

[52] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics—Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[53] J. Rhee, J. Im, G. J. Carbone, and J. R. Jensen, “Delineation of climate regions using in-situ and remotely sensed data for the Carolinas,” *Remote Sensing of Environment*, vol. 112, no. 6, pp. 3099–3111, 2008.

[54] D. E. Stooksbury and P. J. Michaels, “Cluster analysis of southeastern U.S. climate stations,” *Theoretical and Applied Climatology*, vol. 44, no. 3-4, pp. 143–150, 1991.

[55] P. Jaccard, “The distribution of the flora in the alpine zone zone.1,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[56] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification,” *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[57] J. Ghosh and A. Strehl, “Similarity-based text clustering: a comparative study,” in *Grouping Multidimensional Data*, J. Kogan, C. Nicholas, and M. Teboulle, Eds., Springer-Verlag, Berlin, Germany, 2006.

[58] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Berlin, Germany, 2009.

[59] J. M. Peña, J. A. Lozano, and P. Larrañaga, “An empirical comparison of four initialization methods for the *K*-means algorithm,” *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027–1040, 1999.

[60] D. Arthur and S. Vassilvitskii, “*k*-means++: the advantages of careful seeding,” in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, USA, January 2006.

[61] M. E. Celebi and H. A. Kingravi, “Linear, deterministic, and order-invariant initialization methods for the *K*-means clustering algorithm,” in *Partitional Clustering Algorithms*, M. E. Celebi, Ed., Springer International Publishing, Berlin, Germany, 2015.

[62] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, and M. Adjouadi, “A comprehensive survey on impulse and gaussian denoising filters for digital images,” *Signal Processing*, vol. 157, pp. 236–260, 2019.

[63] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.