Journal of Advanced Informatics in Water, Soil, and Structure 2025 2 (2) 279-299

Journal of Advanced Informatics in Water, Soil, and Structure

journal homepage: www.wss.torbath.ac.ir

Research Article

Clustering precipitation data using an adjusted Davies-Bouldin validity index

Reza Rajabi[1*]

[1]Department of Civil Engineering, Sharif University of Technology, Azadi Avenue, Tehran, Iran.
* Corresponding author: Reza Rajabi: reza.rajabi@sharif.edu
OrCID: 0000-0002-8232-3825

ARTICLE INFO

Received date: 06 Jul 2025
Accept date: 06 Nov 2025
Published date: 20 Nov 2025

Keywords:
Clustering; Validity indices; DB index; Climate classification; MENA, Gridded precipitation products

Abstract

Cluster analysis in climate studies can define climatic zones and provide valuable information. Validation of the clustering results can be complex, particularly for datasets with no distinguishable boundaries. A proper validity index should consider two main problems: unnecessary cluster merging and unnecessary cluster division to identify precise boundaries. This study demonstrates the weaknesses of three common validity indices (DB, CS, Silhouette) for clustering validation and develops an adjusted DB index (DB_R). The DB_R index uses the concepts of spectral clustering and the neighborhood network of data points to control unnecessary merging and division. We use this developed index to validate k-means clustering of precipitation data over the Middle East and North Africa (MENA) to find proper clusters. For comparison purposes, six precipitation datasets, CRU, GPCC, PREC/L, UDEL, CPC daily, and CPC monthly, at comparable spatial resolutions, are examined. Results show that six, six, five, four, five, and five clusters delineate the precipitation zones over MENA for the respective datasets. The CRU and GPCC clusters are more consistent with the Köppen-Geiger Classification (KGC) and previous studies. Results indicate that not only the choice of dataset but also the data resolution can affect clustering results. The clustered climate data can be effectively used in adaptation and mitigation planning.

Homepage: www.wss.torbath.ac.ir

*Corresponding Author:
Rajabi Reza
Email: reza.rajabi@sharif.edu
ORCID: 0009-0002-8232-3825

https://doi.org/10.22048/wss.2026.580646.1025

How to cite this article:
Rajabi,R . (2025). Clustering precipitation data using an adjusted Davies-Bouldin validity index. *Journal of Advanced Informatics in Water, Soil, and Structure*, *1*(2), 279-299. doi: 10.22048/wss.2026.580646.1025

1 Introduction

As a fundamental part of the Earth's water cycle, precipitation significantly influences freshwater availability, environmental quality, and agricultural productivity. Changes in rainfall patterns, driven by global climate shifts, have become a serious concern worldwide. Both excessive and insufficient rainfall can trigger catastrophic events such as floods and droughts. For instance, according to reports from National Geographic, annual global losses from floods exceed $40 billion (Hunt and Menon 2020). Similarly, heavy rainfall in the United States leads to yearly losses of about $8 billion and around 100 deaths annually, with a rising trend. Beyond immediate destruction, extreme dry spells also affect public health, ecosystems, financial systems, and social stability.

To mitigate these climate-related risks, such as rising temperatures and shifting rainfall patterns, thorough data analysis using observations, models, and satellite records is essential. Detecting early warning signs through data processing and pattern recognition can help anticipate potential threats. Developing such pattern recognition systems often involves training models using either supervised or unsupervised learning. Supervised learning uses labeled input-output pairs to identify relationships between variables (Moghim and Bras 2019). In contrast, unsupervised learning identifies hidden structures in unlabeled data. Clustering, a core unsupervised technique, groups data points so that items within the same cluster are more similar to each other than to those in separate clusters.

Clustering has broad applications across disciplines such as biology (Subhani et al. 2010), climatology (Steinbach et al. 2003), medicine (Wismüller et al. 2002), finance (Guam and Jiang 2007), environmental science (Scotto et al. 2010), energy (Iglesias and Kastner 2013), psychology (Kurbalija et al. 2012), and astronomy (Rebbapragada et al. 2009). The effectiveness of clustering depends on three key components: the chosen similarity or dissimilarity measure, the clustering algorithm, and the validation method. Selecting an appropriate similarity measure remains difficult, as it must reflect data values, features, and patterns appropriately (Liao 2005; Aghabozorgi et al. 2015). This measure is then used in algorithms ranging from basic methods like k-means and hierarchical clustering to advanced ones like Self-Organizing Maps (Kohonen 1990), DBSCAN (Ester et al. 1996), and metaheuristic approaches (Das et al. 2007; Liu et al. 2011). Finally, validation uses internal or external indices (Aghabozorgi et al. 2015). Where true labels are known, external indices apply; in fields like climate science, where labels are often uncertain, internal indices are relied upon.

Unlike traditional climate classification based on heuristic rules, clustering offers a fully data-driven way to define climate zones (Zscheischler et al. 2012). For example, Fovell (1997) defined U.S. climate zones by separately clustering monthly temperature and precipitation time series from 343 regions between 1931 and 1981 using Euclidean distance. The results showed that analyzing variables separately offers clearer insights. Lund and Li (2009) identified Colorado climate zones using three hierarchical algorithms applied to temperature data from 292 stations, introducing an ARMA-based similarity measure that accounts for both mean and autocorrelation. Netzel and Stepinski (2016) compared clustering outputs with the Köppen–Geiger system and determined that precipitation, temperature, and temperature range are the most critical variables, with clustering yielding more internally homogeneous and externally distinct classes. Other relevant studies include Michaelides et al. (2001), who used SOM on Cyprus rainfall data and found it more robust than hierarchical clustering, as well as work by Fovell and Fovell (1993), Unal et al. (2003), and Corporal-Lodangco and Leslie (2017).

Despite the power of clustering for discovering hidden data structures, identifying the optimal number of clusters remains a persistent challenge (Jain et al. 1999; Kalton et al. 2001; Dalton et al. 2009). This difficulty is especially pronounced in large regions with diverse weather and climate patterns, such as the Middle East and North Africa (MENA), where climate zones lack sharp boundaries. To address the limitations of conventional validity indices, such as the Davies–Bouldin (DB), Calinski–Harabasz (CS), and Silhouette indices, which often fail to accurately evaluate clustering results in datasets without clearly defined boundaries, this study develops a modified index that incorporates concepts from spectral clustering and neighborhood networks to control two common issues: unnecessary merging

and unnecessary division of clusters. Unlike traditional indices that rely solely on distance-based compactness, the new index considers the local connectivity and shared neighbors among data points. This allows for more precise identification of cluster boundaries, particularly in complex spatial datasets such as climate data. The purpose of the index is to provide a more robust internal validation tool for clustering algorithms like k-means, especially when applied to precipitation data over regions with gradual transitions between climatic zones. We apply clustering to six gridded precipitation datasets and assess them using the proposed index. Sections 2 and 3 describe the study area and datasets, Section 4 outlines the methodology, and Sections 5 and 6 present the results and discussion.

2 Study domain

The study domain is MENA, encompassing the Middle East and North Africa, extending from the latitude. 12°8′41″N to 42°6′35″N and longitude from 13°10′29″W to 63°19′48″E (see Fig. 1), which is rich in natural gas and oil reserves. MENA has no standardized definition; in most cases, the MENA region includes Algeria, Bahrain, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Syria, Tunisia, the United Arab Emirates, and Yemen. This region has suffered from natural disasters such as torrential rains resulting from the Active Red Sea Trough (a low-level pressure trough that extends from the African Monsoon over equatorial Africa, northward over the Red Sea region toward the eastern Mediterranean), which affects the Levant, comprising eastern Egypt, the Sinai Peninsula, Israel, Jordan, Lebanon, and Syria. (de Vries et al., 2013). This region is also affected by more frequent and severe droughts; NASA states that the recent drought, which began in 1998 in the eastern Mediterranean Levant region, is likely the worst drought of the past nine centuries. Clustering based on precipitation patterns can reveal varied physical features and precipitation mechanisms across the MENA region. To capture the diversity of precipitation zones more comprehensively, Armenia, Azerbaijan, Turkey, and Cyprus are also included in the study domain (see Fig. 1).

3 Datasets

This study develops an index to improve the efficiency of clustering performance. To demonstrate the effectiveness of this index for precipitation clustering, we use several gridded datasets over the study domain, as follows.

3.1 Climatic Research Unit (CRU)

Monthly time series of 10 climatic variables, including cloud cover, diurnal temperature range, frost day frequency, potential evapotranspiration (PET), precipitation, daily mean temperature, monthly average daily maximum and minimum temperature, and vapor pressure on 0.5×0.5 degree spatial resolution is provided by the CRU at the University of East Anglia (Harris & Jones, 2017). CRU dataset is commonly used in climate studies. For instance, Zhao and Fu (2006) compared products from ERA-40, NCEP-2, and CRU

for summer precipitation over China and found that the CRU data and observation station data match well. This study uses monthly precipitation from CRU TS 4.01 with temporal coverage of 1901-2016 as the first dataset (data are publicly available at http://archive.ceda.ac.uk/).

3.2 Global Precipitation Climatology Center (GPCC)

The Global Precipitation Climatology Center (GPCC) was established in 1989 at the request of the World Meteorological Organization (WMO), and is operated by the Deutscher Wetterdienst (DWD, National Meteorological Service of Germany) as a German contribution to the World Climate Research Programme (WCRP) (Becker et al., 2011). GPCC provides global monthly precipitation using observations from 65335 stations over the world. The main advantage of GPCC is the number of observation stations used, which is considerably greater than in other datasets. (Becker et al., 2011; Sun et al., 2018). GPCC products are widely used in climate studies, such as the validation of satellite rainfall datasets (Nicholson et al., 2003), evaluating the hydrological cycle (Schneider et al., 2017), drought monitoring (Raziei et al., 2011), and climate classification (Kottek et al., 2006). The Full Data Reanalysis v7 product of GPCC with temporal coverage of 1891-2016 on 0.5×0.5 degree geographic resolution is the second selected dataset for this study (data are publicly available at https://www.dwd.de or https://psl.noaa.gov/).

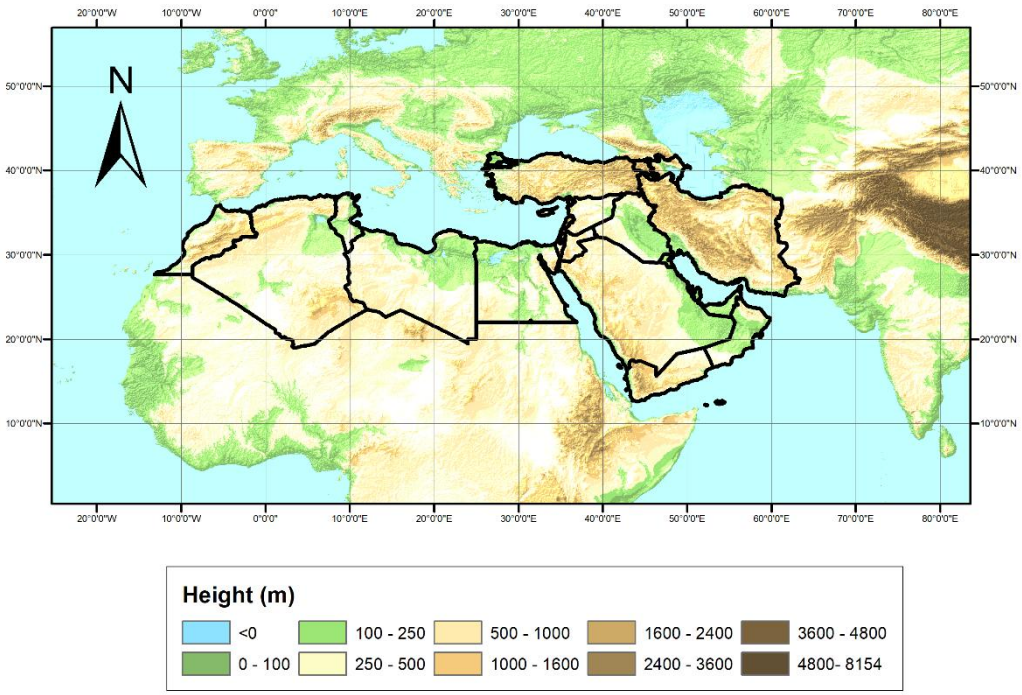

Fig. 1 Study domain including the Middle East and North Africa

3.3 PREC/L

The PRECipitation REConstruction (PREC) dataset, established at the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center, provides precipitation data over both oceans and land. The PREC/L dataset is the land version of PREC, which interpolates gauge observations from more than 17,000 stations collected in the Global Historical Climatology Network (GHCN) Version 2 and the Climate Anomaly Monitoring System, CAMS datasets (Chen et al., 2002). This dataset can be used for temporal and spatial assessment of precipitation. (Zhang et al., 2012). This study uses the monthly gridded PREC/L on 0.5×0.5 degree resolution and temporal coverage of 1948-2011 (data are publicly available at ftp://ftp.cpc.ncep.noaa.gov/precip/50yr/ or https://psl.noaa.gov/).

3.4 UDEL

UDEL is a dataset from the University of Delaware, which provides monthly precipitation and temperature time series at 0.5-degree spatial resolution over land0.5×0.5 (Willmott, 2000). Like the aforementioned datasets, UDEL can serve as a comparative dataset in climate studies. (McAfee et al., 2014). This study uses version V5.01 of UDEL during 1900-2017 as the fourth dataset for precipitation clustering over MENA. Data are available at http://climate.geog.udel.edu/ or https://psl.noaa.gov/.

3.5 Climate Project Center (CPC)

The CPC unified gauge-based analysis of global daily precipitation is a new dataset provided by the NOAA Climate Prediction Center (CPC). Gauge information from more than 30,000 stations from multiple sources, including WMO's Global Telecommunication System (GTS), Cooperative Observer Program (COOP), and other national and international agencies, is interpolated to produce an improved precipitation dataset. (Xie et al., 2010). This study uses the long-term mean daily precipitation of CPC (hereafter called CDCD) from 1979 to the present as the fifth dataset (data are publicly available at https://psl.noaa.gov/). Also, we use monthly precipitation by averaging daily CPC (hereafter called CDCM).

Annual precipitation time series of the datasets used over the study domain are shown in Fig. 2. Although all datasets show similar trends, UDEL precipitation is overestimated at the beginning of the period, and CPC precipitation is underestimated at the end of the period. Differences between these datasets can arise from the varied observation stations used or from different interpolation methods.

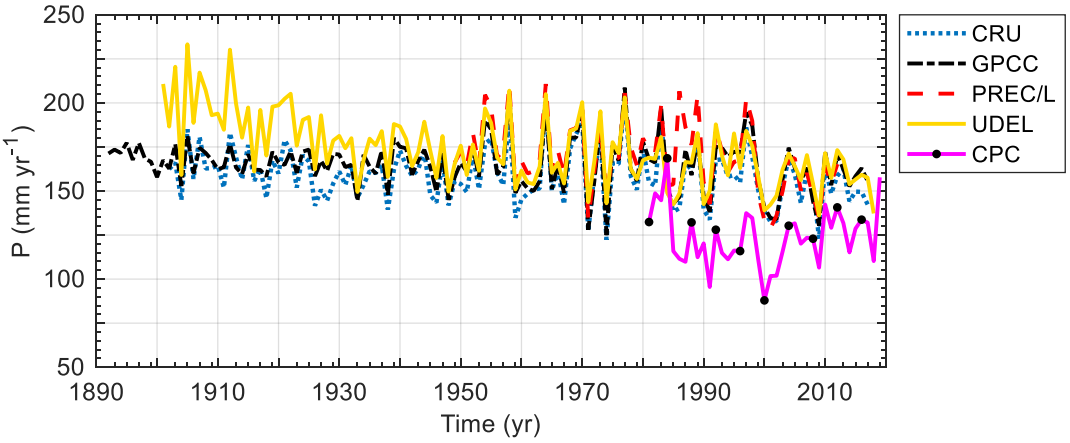

Fig. 2 Annual precipitation time series of the used datasets over the MENA

4 Methodology

This paper aims to introduce a new internal validity index for complex datasets such as climatic ones. Prior to clustering validation, the similarity/dissimilarity measure and clustering algorithm should be determined. One of the most commonly used measures for similarity/dissimilarity assessment is the Minkowski distance, as (Lhermitte et al., 2011; Liao, 2005; Serra & Arcos, 2014)

$$d_{xy} = \left(\sum_{i=1}^{p} \left(x_i - y_i \right)^t \right)^{\frac{1}{t}} \quad (1)$$

where d_{xy} is the Minkowski distance of two p-dimensional data points (x,y) and t is a parameter. This study uses $t = 2$, which results in Euclidean distance as the selected similarity/dissimilarity measure. To produce clusters based on the similarity/dissimilarity matrix, we use k-means, the most popular clustering algorithm. Although k-means is commonly used due to its simplicity and high convergence speed, cluster centroid initialization remains challenging. Standard k-means initializes cluster centroids randomly and refines them iteratively. This random initialization can lead to local optima (Tan et al., 2006). To avoid local optima, we run the k-means algorithm 200 times for each dataset to find the best solution through the validation process.

To evaluate clustering results and determine the optimal number of clusters, validity indices are required. Various factors, including compactness, cluster density, closeness, separation, and between-cluster distances, must be combined to formulate validity indices. Three traditional validity indices, including DB, CS, and Silhouette indices, are used for comparison purposes to evaluate the efficiency of our developed validity index. One of the most popular validity indices is the DB index. (Davies & Bouldin, 1979). All parameters are defined upon their first appearance in the order of equations. Any parameter not redefined in subsequent equations retains its earlier definition.

$$\mathrm{DB}(c) = \frac{1}{c} \sum_{i=1}^{c} R_i \quad (2)$$

$$R_i = \max_{j=1,\ldots,c} \frac{S_i + S_j}{D_{ij}} \quad (3)$$

where c is the number of clusters, S is the compactness measure corresponding to clusters i and j, and D_{ij} is the separation measure between clusters i and j (j = 1,2, …, c). The compactness and separation measures are calculated as

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} d\left(x, \bar{C}_i \right) \quad (4)$$

$$D_{ij} = d\left(\bar{C}_i, \bar{C}_j \right) \quad (5)$$

where x indicates data points, $(\bar{C}_i, \bar{C}_j)$ are centroids of clusters i and j, and d is the similarity/dissimilarity function. The value of c that minimizes the DB index can be regarded as the optimal number of clusters. Another validity index is the CS index proposed by Chou et al. (2004), which computes the compactness of the clusters based on the maximum distance between the data points within them, and the separation measure is calculated based on the minimum distance between cluster centroids. Similar to the DB index, the value of c yielding the minimum CS index corresponds to the optimal number of clusters. The Silhouette (Sil) index, another validity index, uses distances between data points in the same cluster for the compactness measure and the nearest neighbor distance for the separation measure (Arbelaitz et al., 2013; Rousseeuw, 1987). The optimal value of c is obtained by maximizing the Silhouette index.

As the motivation for this study, we elucidate the weaknesses of traditional validity indices in determining the optimal number of clusters for climate datasets such as precipitation. First, to demonstrate the complexity of the clustering approach for climate studies, we consider each pixel of the study domain as a data point. Then, we assign the long-term monthly precipitation of that pixel to the corresponding data point as its features. We obtain a set of data points with 12 dimensions corresponding to months of the year. Then, we use the Principal Component Analysis (PCA) on these datasets to reduce dimension of the data for results visualization. Results of the PCA show that the two first principal components related to the first largest eigenvalues (P_1 and P_2 in Fig. 3) cover about 92.19, 92.15, 87.53, 88.84, and 93.73 percent of the total variance of long-term monthly precipitation of CRU, GPCC, PREC, UDEL, and CPC, respectively. These high values of variance coverage ensure that the first two principal components of the PCA are reliable indicators for representing the overall cluster positions (Fig. 3).

Fig. 3 shows that there is no clear cluster boundary (distinguished clusters) in datasets, which reveals the inability of clustering validation methods to determine the optimal number of clusters. This work develops a new validity index capable of identifying proper clusters for complex datasets that lack explicit cluster boundaries. To determine the optimal number of clusters and specify the best possible cluster boundaries in such complex datasets, a robust validity index is needed that can prevent both unnecessary cluster merging and unnecessary cluster division, which are two main problems in clustering validation (Kim & Ramakrishna, 2005; Žalik & Žalik, 2011). To further examine these problems, three main regions of the experimental data (R_1, R_2, and R_3) are shown in Fig. 4.

Table 1 indicates that the three optimal clusters (shown with different colors and symbols) in subregion R_{11} can be detected by all three validity indices. Thus, using validity indices in each individual region can lead to a total of seven clusters in the experimental data. On the other hand, we used the validity indices for clustering validation of all regions (considered as one domain) of the experimental data. Results of the clustering validation of this whole experimental domain are summarized in Table 2.

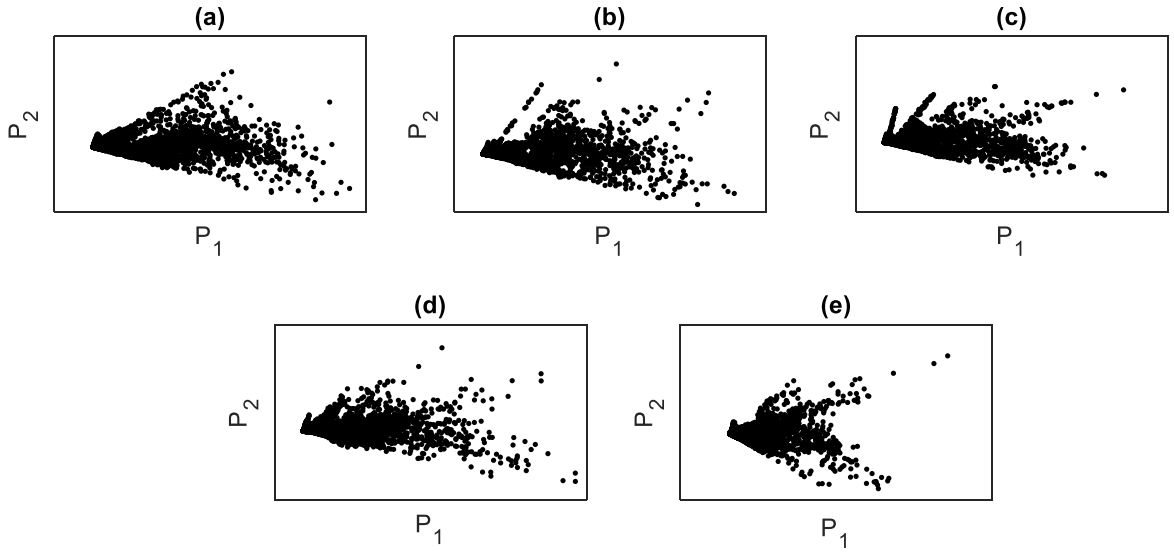

Fig. 3 Results of the PCA long-term monthly precipitation for (a) CRU, (b) GPCC, (c) PREC/L, (d) UDEL, and (e) CPC, where P1 and P2 are the two first principal components corresponding to the two first largest eigenvalues..

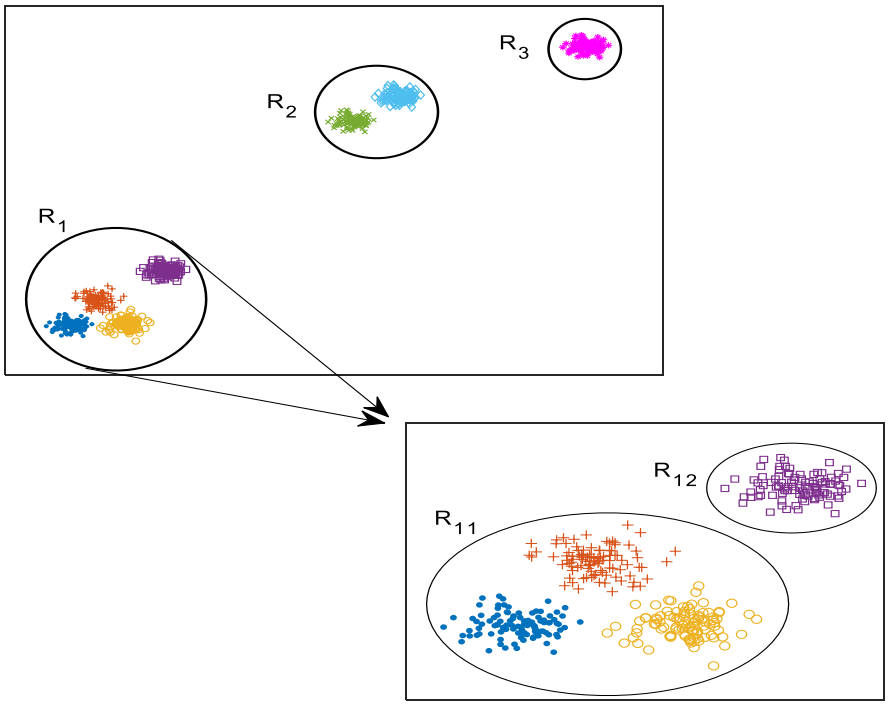

Fig. 4 Experimental data to examine the unnecessary cluster merging problem

Table 1 Clustering validation results for region R_{11} of the experimental data

Index	Number of Clusters								
	2	3	4	5	6	7	8	9	10
DB	0.825	0.424	0.762	0.934	0.904	0.984	0.946	0.910	0.887
Sil	0.592	0.862	0.733	0.618	0.610	0.529	0.531	0.530	0.521
CS	1.225	0.683	0.979	0.966	1.039	1.229	1.384	1.287	1.349

Table 2 indicates that the optimal number of clusters for the experimental data is equal to three based on the DB, Sil, and CS indices. This result can reveal that the validity indices are not able to recognize the internal clusters of the regions (e.g., 2 in R_2 and 4 in R_1) when the regions are considered together (as one domain of data). This example can show unnecessary cluster merging. Thus, for a robust clustering analysis, more details about the structure of the dataset (e.g., connectivity between data points) in clustering validation are required.

Table 2 Validation of the k-means clustering on the experimental data by DB, Sil, and CS indices

Index	Number of Clusters								
	2	3	4	5	6	7	8	9	10
DB	0.302	0.190	0.300	0.324	0.468	0.334	0.499	0.635	0.764
Sil	0.942	0.966	0.889	0.880	0.797	0.912	0.856	0.807	0.748
CS	0.392	0.280	0.318	0.398	0.366	0.419	0.476	0.489	0.558

This work uses spectral clustering to consider the structure of the data. Spectral clustering, which is used in image segmentation (Zeng et al., 2014), speech separation (Bach & Jordan, 2006), and even climate studies (Türkeş & Tatlı, 2011), combines the concept of clustering and graph theory to cluster datasets of different shapes and densities (Ng et al., 2002; Von Luxburg, 2007). The spectral clustering algorithm can apply a Gaussian kernel function to define connectivity between vertices (data points) of the graph as

$$A_{xy} = \exp\left(-\frac{d_{xy}^{\,2}}{2\sigma^2}\right) \quad (6)$$

where d_{xy} is the Euclidean distance between data points x, y, and σ is the kernel function parameter to control the width of neighborhood. The greater value of A_{xy} indicates the stronger connectivity between points x and y.

To further illustrate the merging clusters issue, three points including “a”, “b”, and “c” are considered in Fig. 5. Although the distance from point “a” to point “b” is equal to distance from point “c” to point “b”, point “a” belongs to a different cluster, Cluster 1, while points “b” and “c” belong to Cluster 2 (see in Fig. 5). Results reveal that additional considerations are necessary to make the data points within the same cluster more similar to each other and thus internal clusters (e.g. in Fig. 4) can be distinguished and unnecessary cluster merging can be controlled.

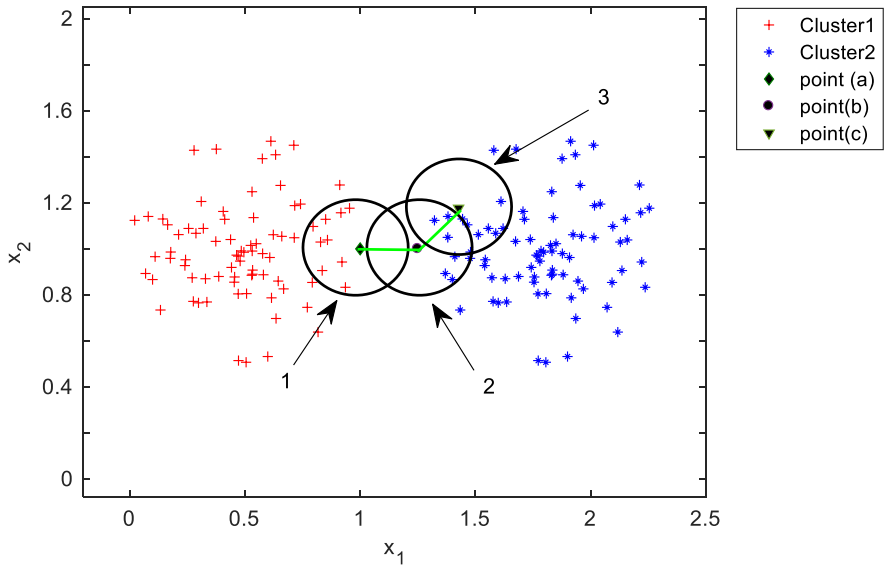

Fig. 5 A test case for the merging clusters issue. Circles 1, 2, and 3 are "ε" -neighborhood of points “a”, “b”, and “c”.

Zhang et al. (2011) proposed a revised Gaussian kernel function to monitor the unnecessary merging problem using the Common Nearest Neighbors (CNN) measure. They defined CNN as the number of common neighbors located in ε-neighborhood of the two points. A similar measure called Shared Nearest Neighbors (SNN) calculates the number of shared neighbors of the two points among k-nearest neighbors instead of using parameter ε (He et al., 2015; Yuan et al., 2016). To have a visual understanding of the CNN, consider circles 1, 2, and 3 in ***Error! Reference source not found.*** as the ε-neighborhood of points “a”, “b”, and “c”, respectively. We can see that circles 1 and 2 have no common points (CNN = 0), while circles 2 and 3 have three common neighbors (CNN =3). Thus, the Gaussian kernel function can be modified to make points in the same cluster closer to each other since they have more common neighbors compared to the other points from different clusters, as

$$A_{xy} = \exp\left(-\frac{{d_{xy}}^2}{2\sigma^2\left(1+\text{CNN or SNN}\right)}\right) \qquad (7)$$

All parameters are similar to those in Eq. 6. This equation, which is the modified form of the Gaussian kernel function, can take into account the neighborhood network of the dataset to make the connectivity of the points within the same cluster stronger. We use this modification to revise the compactness measure of the DB index to make it more powerful in detecting unnecessary merging problems. In the revised DB index, the similarity/dissimilarity function (d in Eq. 4) is replaced by the new Gaussian kernel function. But an important point should be taken into account. A greater value of d in Eq. 1 and Eq. 4 leads to more dissimilar data points and less compact clusters, respectively, while due to the negative sign in Eq. 7, a greater value of d leads to a smaller value of A_{xy}, which means weaker connectivity between x and y. Indeed, d_{xy} and A_{xy} have the same meaning implicitly. In other words, a greater value of d_{xy} refers to more dissimilar data points (x and y) with weaker connectivity (smaller A_{xy}) in the graph of spectral clustering. To have the same quantitative form of Eq. 1 and Eq. 7, we modify A_{xy} as

$$(A_{\mathrm{R}})_{xy} = 1-\exp\left(-\frac{{d_{xy}}^2}{2\sigma^2\left(1+\text{CNN or SNN}\right)}\right) \qquad (8)$$

where the value of $(A_{\mathrm{R}})_{xy}$ varies between zero for totally different data points and one for the two well-matched ones. By this modification, we

transformed a connectivity function in the graph (Eq. 7) to a similarity/dissimilarity function, which is required for the compactness measure in Eq. 4. The revised compactness measure of cluster *i*, $(S_{\mathrm{R}})_i$, can be expressed as

$$(S_{\mathrm{R}})_i = \frac{1}{|C_i|}\sum_{x\in C_i}\underset{y\in C_i}{\mathrm{mean}}\left(A_{\mathrm{R}}\right)_{xy} \quad (9)$$

All variables are explained in Eqs. 4 and 8. Indeed, this compactness measure computes compactness of a target cluster based on the average distance between data points within that cluster, while Eq. 4 computes this value based on the average distance between data points of a cluster and the cluster centroid of the corresponding cluster.

The points located in different clusters should have the smallest possible number of common neighbors with each other. In other words, the summation of the measures of CNN or SNN for all data points in different clusters should be minimum, which can be used to control unnecessary division. When unnecessary cluster division occurs, data points belonging to one cluster can be wrongly divided into two clusters, and thus the summation of CNN or SNN measures between data points of the two resulting clusters will be large. We can use the number of sharing neighbors between points of those unnecessary clusters to make their centroids closer. A smaller separation measure in Eq. 3 leads to a greater DB index, which is the indicator of a non-optimal clustering result. Thus, the new separation measure for the revised DB index can be expressed as

$$(D_{\mathrm{R}})_{ij} = 1-\exp\left(-\frac{d\left(\bar{C}_i,\bar{C}_j\right)}{2\sigma^2\left(1+\sum_{x\in C_i}\sum_{y\in C_j}\mathrm{CNN\ or\ SNN}(x,y)\right)}\right) \quad (10)$$

where $(D_{\mathrm{R}})_{ij}$ is the revised separation measure and all variables are similar to Eq.5 and Eq. 8. The kernel function parameter corresponding to the local kernel function parameter at point *i* (σ_i) can be considered as the distance between point *i* and the k^{th} nearest neighbor. Zelnik-Manor and Perona (2005) suggested $k=\log(n)$, where *n* is the total number of data points in the dataset. This study uses the average of σ_i for all data points as the kernel function parameter, where σ_i is considered as the distance between data point *i* and the furthest neighbor obtained by the NC algorithm. For instance, the parameter of the kernel function for the experimental data in Fig. 4 is equal to 0.447 ($\sigma=0.447$). Equation 10 confirms that a greater value of CNN or SNN leads to closer cluster centroids. Thus, the new validity index (DB_R) can be obtained from Eq. 2 using the new compactness (Eq. 9) and separation (Eq. 10) measures.

Although using CNN or SNN measures is helpful to control the two main discussed problems of the clustering validation process, assigning proper values to ε in CNN or *k* in SNN is challenging. This study uses the Neighborhood Construction (NC) algorithm to find proper neighbors of the data points based on the connectivity, density, and local features of the dataset (İnkaya et al., 2015). To evaluate the skill of the new validity index (DB_R), this revised DB index (see Eq. 4) with new compactness and separation measures is used to cluster the test data in Fig. 4. Results are summarized in Table 3.

Results show that the seven-cluster structure leads to the minimum DB_R and thus avoids merging issue. In addition, DB_R increases noticeably in going from seven to 8 and more clusters, which indicates unnecessary cluster division; while this significant change is not evident in values of the DB, CS, and Silhouette indices (Table 2).

Table 3 Results of the clustering of the test data (Fig. 4) based on the new validity criteria (DB_R)

Number of clusters	2	3	4	5	6	7	8	9	10
DBR	1.930	1.892	1.853	1.842	1.820	1.766	2.505	3.183	3.361

5 Results

The ability of the developed validity index (DB_R) to cluster datasets with no clear cluster boundaries is evaluated using six precipitation datasets over MENA (Section 3). To demonstrate the effectiveness of DB_R, validation results are compared using the DB, CS, Silhouette, and DB_R indices in Fig. 6, and the optimal number of clusters for each dataset according to these validity indices is summarized in Table 4.

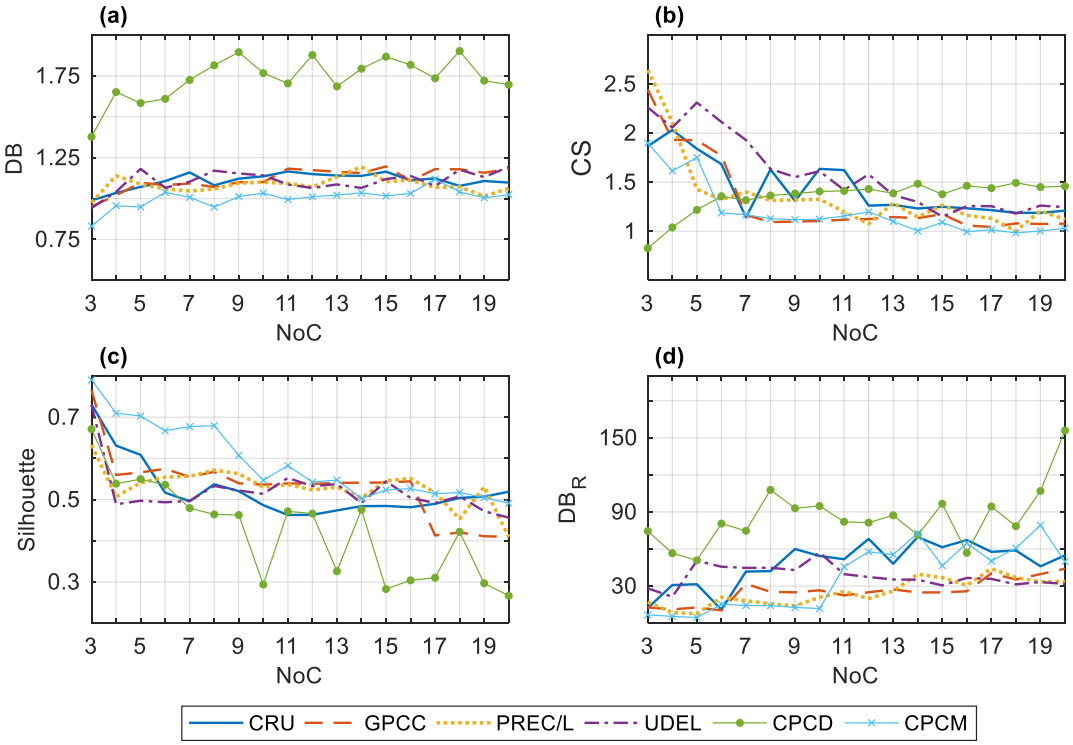

Fig. 6 Clustering validation for the used precipitation datasets including CRU, GPCC, PREC/L, UDEL, CPCD, and CPCM based on the (a) DB index, (b) CS index, (c) Silhouette index (Sil), and (d) the developed validity index (DB_R) where NoC is the number of clusters, ranging from 3 to 20.

For all datasets, clustering validation using the DB and Sil indices yielded three clusters as the optimal number. Although clustering validation based on the CS index for the experimental data (Table 2) showed an unnecessary clustering merging problem, the optimal number of clusters based on the CS index is significantly larger than those based on other validity indices for all datasets except CPCD. The revised DB index (DB_R), which properly validated the clustering of the test data (Table 3), suggests a greater number of clusters than the DB or Sil indices, thereby controlling the unnecessary merging problem. In addition, the number of clusters found by DB_R is smaller than that proposed by the CS index, which helps control the unnecessary cluster division problem. As the most reasonable index, DB_R is used to determine the optimal number of clusters over the study domain (Fig. 7).

To analyze the results, the long-term monthly precipitation (seasonal variation) of the cluster centroids is presented in Fig. 8. In addition, the empirical Cumulative Distribution Function (eCDF) is used to evaluate precipitation distributions for the validated clusters (Fig. 9). To construct eCDFs of the cluster centroids, the Kaplan–Meier estimator (Lawless, 2011) has been used.

Table 4 Optimal number of the clusters based on the validity indices

Index	Dataset					
	CRU	GPCC	PREC/L	UDEL	CPCD	CPCM
DB	3	3	3	3	3	3
CS	19	17	18	15	3	18
Sil	3	3	3	3	3	3
DB_R	6	6	5	4	5	5

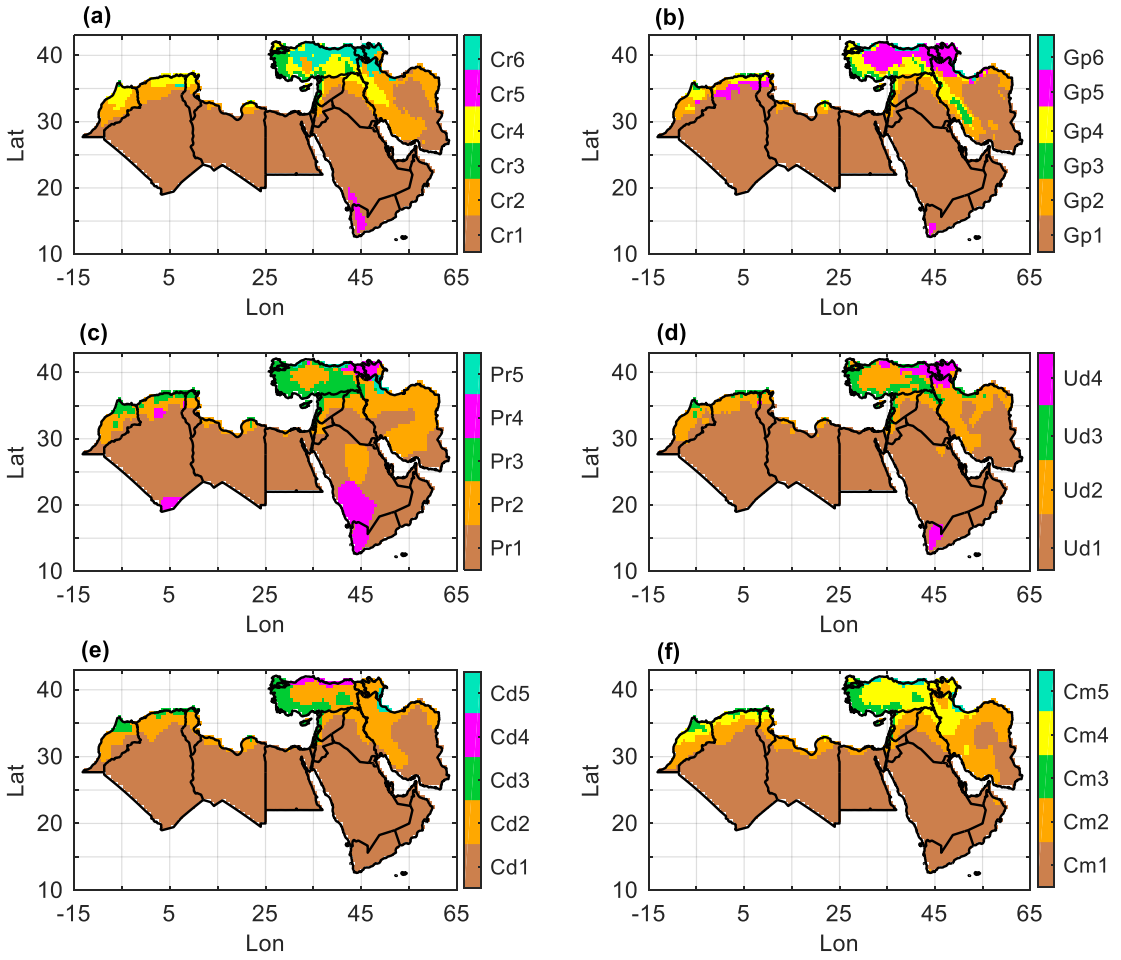

Fig. 7 Geographic distribution of the validated clusters for (a) CRU, Cr; (b) GPCC, Gp; (c) PREC/L, Pr; (d) UDEL, Ud; (e) CPCD, Cd; and (f) CPCM, Cm. Numbers in the colorbar refer to the cluster.

The analysis of the resulting clusters is described for each dataset separately.

5.1 CRU

The developed validity index delineates the study domain (MENA) into six clusters (the optimal number of clusters) based on CRU precipitation (Fig. 7a). The six clusters, including Cr1, Cr2, Cr3, Cr4, Cr5, and Cr6, have experienced long-term mean annual precipitation equal to 60, 274, 715, 510, 337 and 489 mm during 1901-2016, respectively. Also, Fig. 8a shows that cluster 1 (Cr1) is almost dry in all months, while clusters Cr2, Cr3, and Cr4 can be characterized as summer-dry clusters (dry in June, July, August, and September). Clusters Cr5 and Cr6 show different patterns; for instance, Cr5 has its wettest months in summer (July and August), and the wettest month of Cr6 occurs in late spring (May). Fig. 9a shows that clusters Cr1, Cr2, Cr4, and Cr6 have narrower distributions than Cr3 and Cr5. In addition, the distributions of Cr1, Cr4, and Cr6 are more uniform, whereas those of Cr2 and Cr5 are right-skewed. Furthermore, precipitation in Cr3 is approximately normally distributed. Clusters Cr3 and Cr1 exhibit the highest and lowest temporal variability, respectively.

To assess the degree of similarity between data points in a cluster and the cluster centroid, the Degree of Membership (DoM), as defined in Fuzzy C-means clustering, is used (Tan et al., 2006). The DoM varies between 0 and 1, where a larger value indicates greater similarity between a data point and its cluster centroid. To further illustrate the spatial distribution of the clusters, DoM in six CRU clusters is shown in Fig. 10. Data points belonging to a cluster have a larger DoM with respect to that cluster. To be more precise, 95.95%, 90.31%, 87%, 76.89%, 97.29%, and 77.83% of the data points belonging to clusters Cr1, Cr2, Cr3, Cr4, Cr5, and Cr6 have DoM larger than 0.5, respectively.

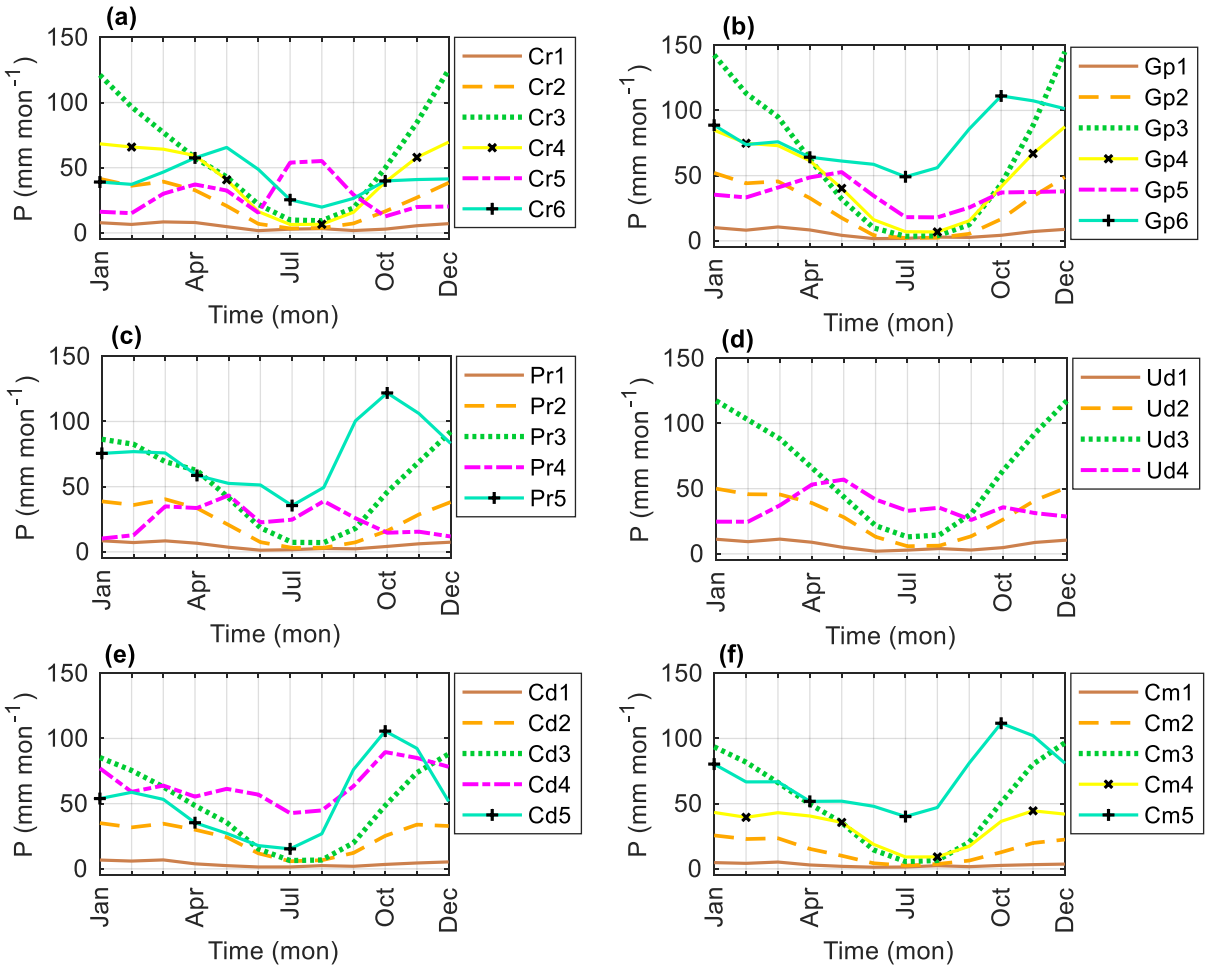

Fig. 8 Seasonal change of precipitation for the cluster centroids of the datasets, including (a) CRU, Cr; (b) GPCC, Gp; (c) PREC/L, Pr; (d) UDEL, Ud; (e) CPCD, Cd; and (f) CPCM, Cm. Numbers in the legend refer to the cluster.

5.2 GPCC

The DB_R index identifies six optimal clusters for GPCC precipitation over MENA (Fig. 7b). Fig. 8b shows that precipitation tends to increase in fall and then decrease to its minimum in summer for the Gp2, Gp3, and Gp4 clusters, while Gp5 reaches its maximum in spring (May). Gp6 exhibits a distinct seasonal precipitation pattern, with a maximum in October followed by a decrease. Although Gp3 has the largest extreme values (Fig. 9b), Gp5 can be considered the wettest cluster, with an average of 934 mm of precipitation per year, followed by Gp3, Gp4, Gp6, Gp2, and Gp1 with 752, 575, 420, 305, and 71mm average annual precipitation. The empirical cumulative distribution functions (eCDFs) of the GPCC cluster centroids show that Gp3, Gp4, and Gp5 have wider distributions (e.g., larger extremes) than the other clusters (see Fig. 9b). In addition, the distributions of Gp2, Gp3, and Gp4 are right-skewed, whereas the precipitation distributions of Gp5 and Gp6 are approximately normal. The lowest and highest variability occur in Gp1 and Gp3, respectively. Degree of membership (DoM) for the GPCC clusters is shown in Fig. 11, where 96.61% of Gp1 cluster members have DoM larger than 0.5, followed by 89.85%, 84.04%, 79.26%, 81.94%, and 75.60% DoM for the members of Gp2, Gp3, Gp4, Gp5, and Gp6 with respect to their corresponding clusters.

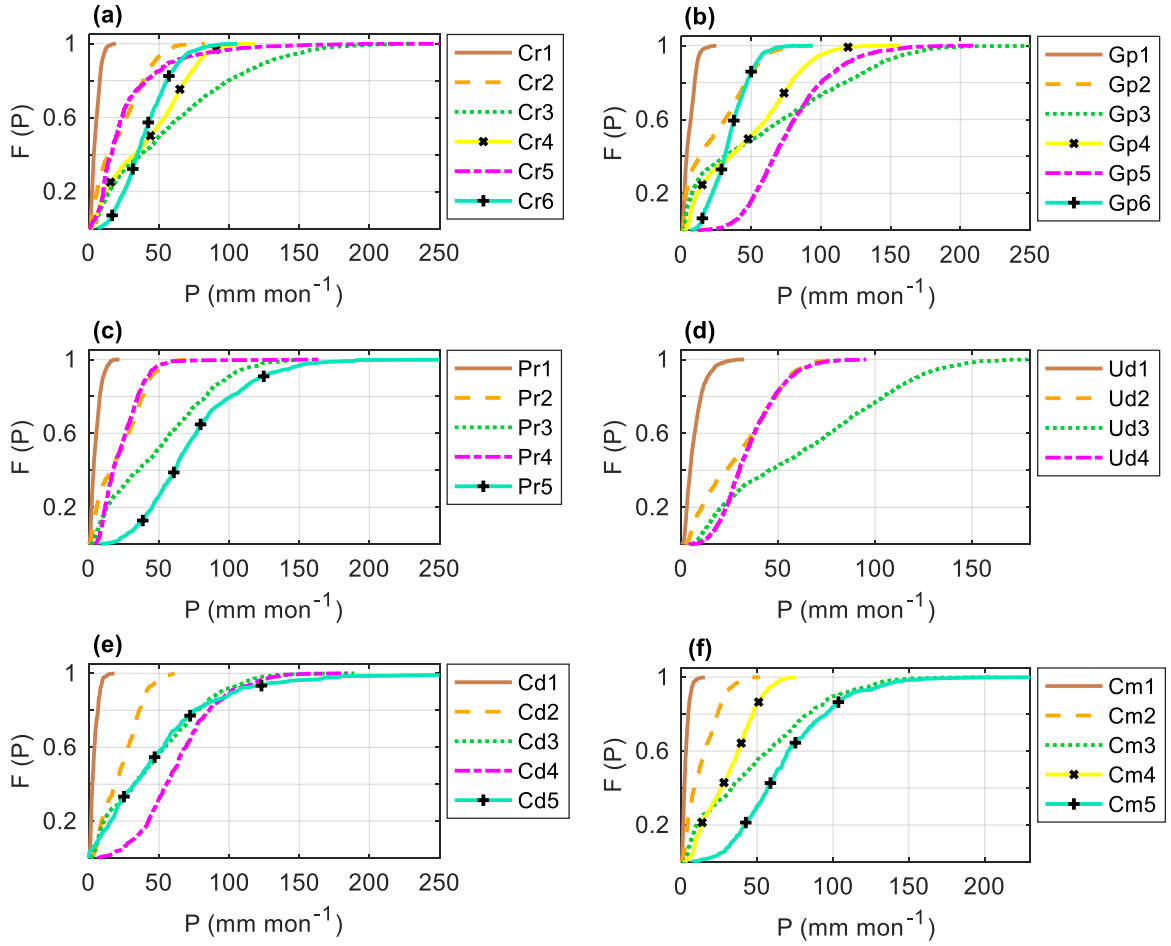

Fig. 9 The precipitation empirical Cumulative Distribution Function (eCDF) of the cluster centroids for the datasets including (a) CRU, Cr; (b) GPCC, Gp; (c) PREC/L, Pr; (d) UDEL, Ud; (e) CPCD, Cd; and (f) CPCM, Cm. Numbers in the legend refer to the cluster.

5.3 PREC/L

The developed index (DB_R) identifies five optimal clusters for PREC/L precipitation over the study domain, with their spatial distribution shown in Fig. 7c. Figure 8c reveals that the seasonal precipitation pattern differs across all clusters. Although the largest precipitation occurs in winter for Pr1, Pr2, and Pr3, the largest precipitation in Pr5 occurs in October. This wettest cluster, with an average of 887 mm per year, has no dry month throughout the year; whereas Pr2 and Pr3, with 273 and 600 mm per year, are nearly dry during the summer months (see Fig. 8c). Cluster Pr4, with its wettest month in spring (May), has an average of 288 mm of precipitation per year. The eCDFs of the cluster centroids show greater variability and more extreme values in Pr5, followed by Pr3 (Fig. 9c). As Fig. 9c shows that clusters Pr1 and Pr2 have narrower and more uniform distributions than the other clusters. By contrast, Pr3, Pr4, and particularly Pr5 experience more extreme precipitation (long-tailed distributions). Clusters Pr5 and Pr1 exhibit the highest and lowest temporal variability, respectively. Although the eCDFs and annual precipitation of clusters Pr2 and Pr4 are similar, their seasonal patterns of long-term monthly precipitation can distinguish these two clusters (see Fig. 8c). The degree of membership for PREC/L clusters is shown in Fig. 12. The strongest similarity to the cluster centroid is found in Pr1, where 98.90% of the cluster members have a DoM greater than 0.5; whereas for clusters Pr2, Pr3, Pr4, and Pr5, 92.37%, 90.85%, 89.04%, and 88.63% of their members, respectively, have a DoM greater than 0.5.

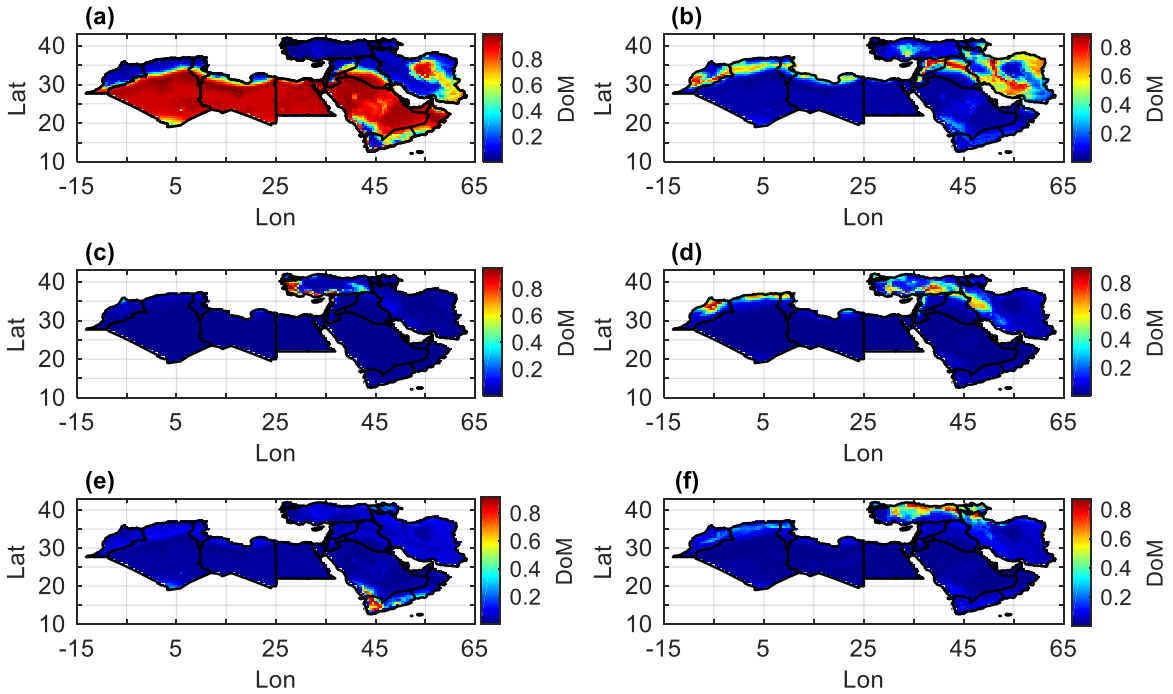

Fig. 10 Degree of Membership of the CRU clusters including (a) Cr1, (b) Cr2, (c) Cr3, (d) Cr4, (e) Cr5, and (f) Cr6

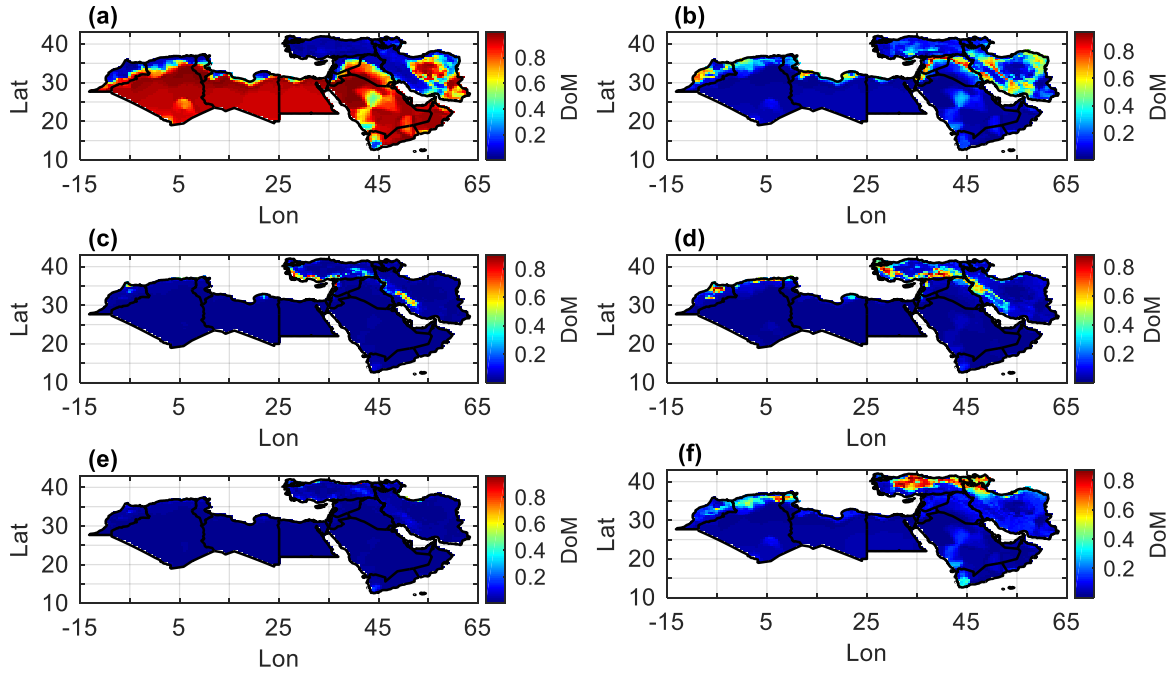

Fig. 11 Degree of Membership of the GPCC clusters including (a) Gp1, (b) Gp2, (c) Gp3, (d) Gp4, (e) Gp5, and (f) Gp6

5.4 UDEL

UDEL precipitation over MENA is clustered into four groups, including Ud1, Ud2, Ud3, and Ud4 with average annual precipitation of 81, 365, 428 and 771 mm, respectively (see Fig. 7d). The seasonal variability of Ud2 and Ud4 is smaller than that of Ud3 (see Fig. 8d). In addition, the seasonal pattern of Ud4 differs from those of Ud2 and Ud3. For instance, Ud4 exhibits smoother precipitation changes, with the largest precipitation in May, while the maximum precipitation for Ud2 and Ud3 occurs in December and January. Fig. 9d shows that clusters Ud1, Ud2, Ud3, and Ud4 have uniform, right-skewed, long-tailed, and normal distributions, respectively. In particular, cluster Ud3 has the widest CDF of precipitation (largest extremes).

Members of the four clusters show strong similarity to their cluster centroids, as confirmed by high DoM values (Fig. 13). Specifically, 99.78%, 94.09%, 91.42%, and 92.30% of the cluster members belonging to Ud1, Ud2, Ud3, and Ud4, respectively, have a DoM greater than 0.5.

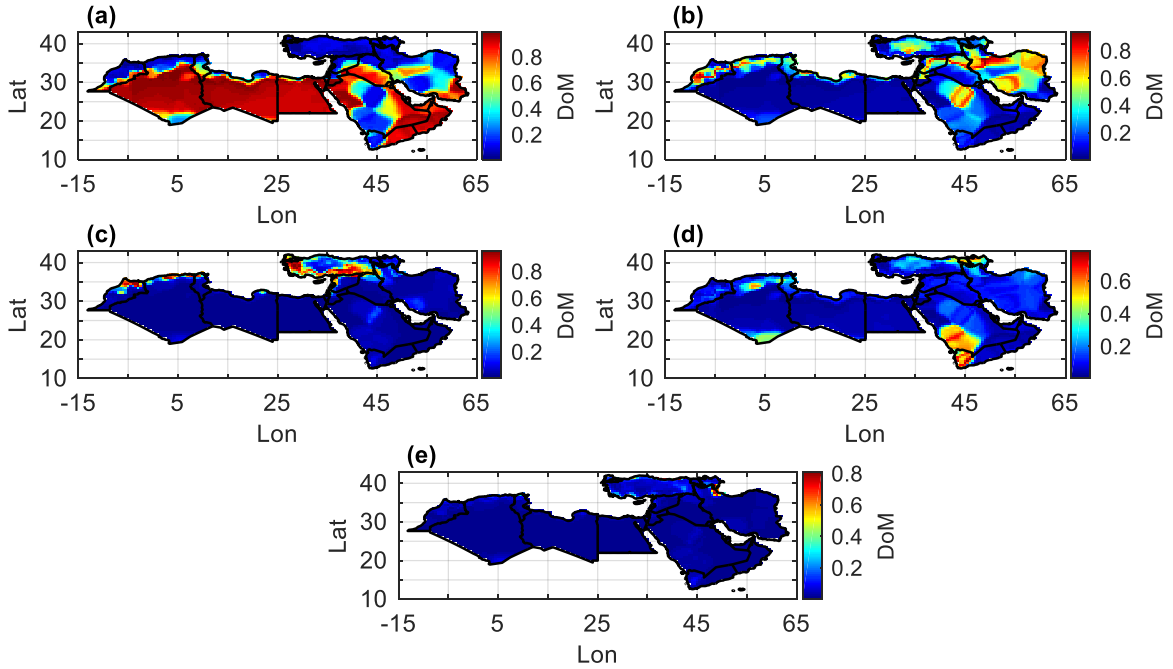

Fig. 12 Degree of Membership of the PREC/L clusters including (a) Pr1, (b) Pr2, (c) Pr3, (d) Pr4, and (e) Pr5

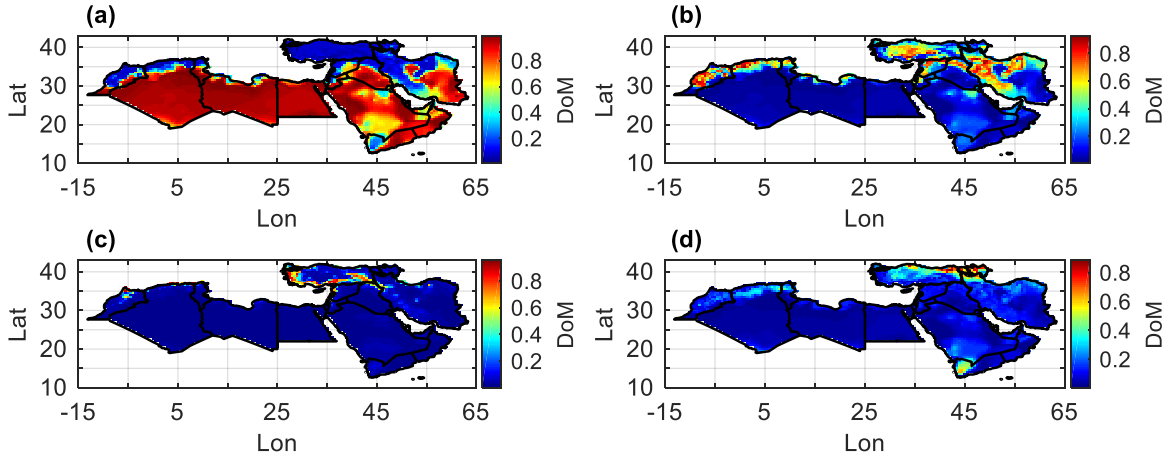

Fig. 13 Degree of Membership of the UDEL clusters, including (a) Ud1, (b) Ud2, (c) Ud3, and (d) Ud4

5.5 CPC

Fig. 7e,f shows that although the newly developed validity index (DB_R) suggests five optimal clusters for both daily and monthly CPC datasets (CPCD and CPCM), the spatial distributions and characteristics (e.g., long-term monthly precipitation and eCDFs) of the clusters obtained from daily and monthly CPC data are significantly different. Clusters Cd1, Cd2, Cd3, Cd4, and Cd5 have experienced mean annual precipitation of 46, 284, 566, 777, and 614 mm, respectively, while these values are 35, 169, 601, 379, and 828 mm for clusters Cm1, Cm2, Cm3, Cm4, and Cm5, respectively. The main difference between the CPCD and CPCM clusters lies in Cd2 and Cm2, where some members of Cd2, located in Turkey, northwestern Iran, and northern African countries, fall into Cm2. The long-term monthly precipitation (Fig. 8e and Fig. 8f) and the eCDFs of monthly precipitation (Fig. 9e and Fig. 9f) show that the CPCM clusters are more distinct than the CPCD clusters. Seasonal precipitation in all CPCD clusters follows a nearly similar pattern, with the largest and smallest variability in Cd5 and Cd2, respectively. In

contrast, the seasonal precipitation patterns of the CPCM clusters differ from one another. For instance, cluster Cd4 has monthly precipitation close to that of Cd3 in the winter months and early spring, and clusters Cd3 and Cd5 have similar CDFs. Clusters Cd3, Cd4, and Cd5 exhibit greater precipitation variability, whereas the largest precipitation variability occurs in Cm3 and Cm5. Although CPCD clusters are obtained using daily CPC directly, CPCM clusters obtained from monthly CPC are more distinct in their daily precipitation CDFs compared to the CPCD ones (see Fig. 14).

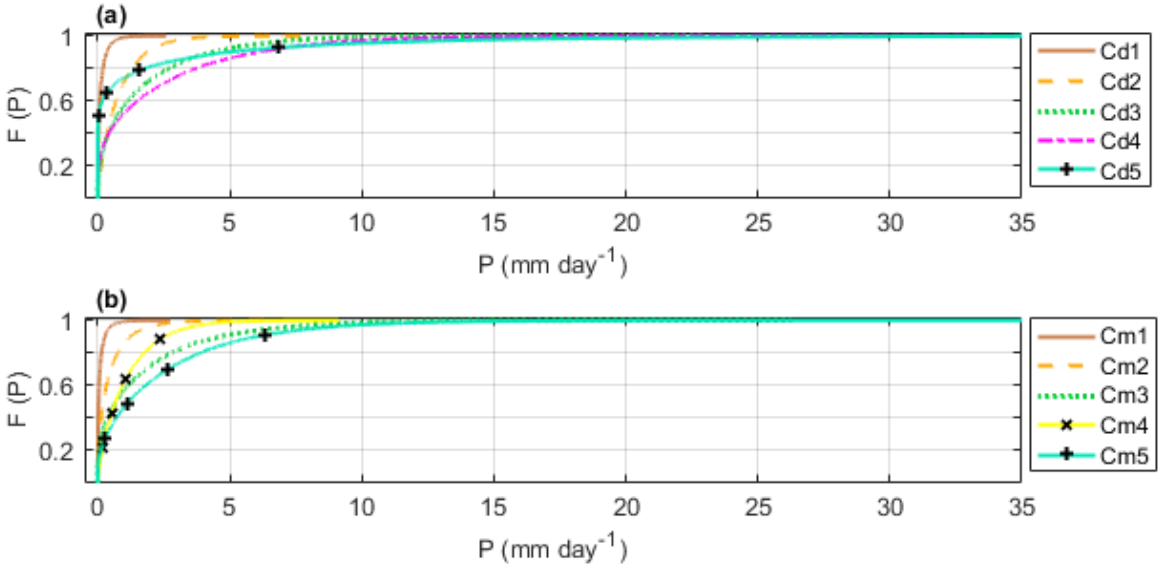

Fig. 14 Daily precipitation eCDFs of the cluster centroids from (a) CPCD and (b) CPCM

To further demonstrate that monthly data is superior to daily data for climate classification, the degree of membership of the CPC clusters is shown in Fig. 15. The results indicate that the DoM values of the CPCM cluster members (100%, 98.72%, 98.64%, 91.54%, and 87.70% related to Cm1, Cm2, Cm3, Cm4, and Cm5 with DoM greater than 0.5) are significantly larger than those of the CPCD cluster members (99.81%, 84.85%, 74.23%, 69.04%, and 65% related to-Cd1, Cd2, Cd3, Cd4, and Cd5 with DoM greater than 0.5). These high DoM values confirm that the CPCM clusters are more compact and reliable than the CPCD clusters.

6 Discussion and Concluding Remarks

The MENA region includes different countries with varied precipitation ranges. Although clustering methods can delineate the region based on precipitation patterns, this knowledge discovery approach requires a proper validity index to determine the optimal number of clusters. This study developed an index called DB_R to validate clustering results. This index is an adjusted version of the DB index, which is widely used in data clustering. To better demonstrate the effectiveness of the developed index, DB_R is compared with other validity indices. The results showed that DB_R accounts for the neighbors of data points to determine the proper number of clusters during the validation process.

To further illustrate the capability of the developed index, this study applied DB_R to evaluate clustering results for precipitation—a highly nonlinear and complex climatic variable—over the MENA region. To account for uncertainties and for comparison purposes, six precipitation datasets were used, including CRU, GPCC, PRECL, UDEL, CPC daily, and CPC monthly. The results showed that although clustering of various datasets can reveal the general characteristics of precipitation over different parts of the region, the number of clusters varies, which may indicate differences in precipitation patterns or distributions. In other words, each dataset can yield a different clustering structure. For instance, the larger number of clusters in the CRU and GPCC datasets can reveal more detailed precipitation patterns.

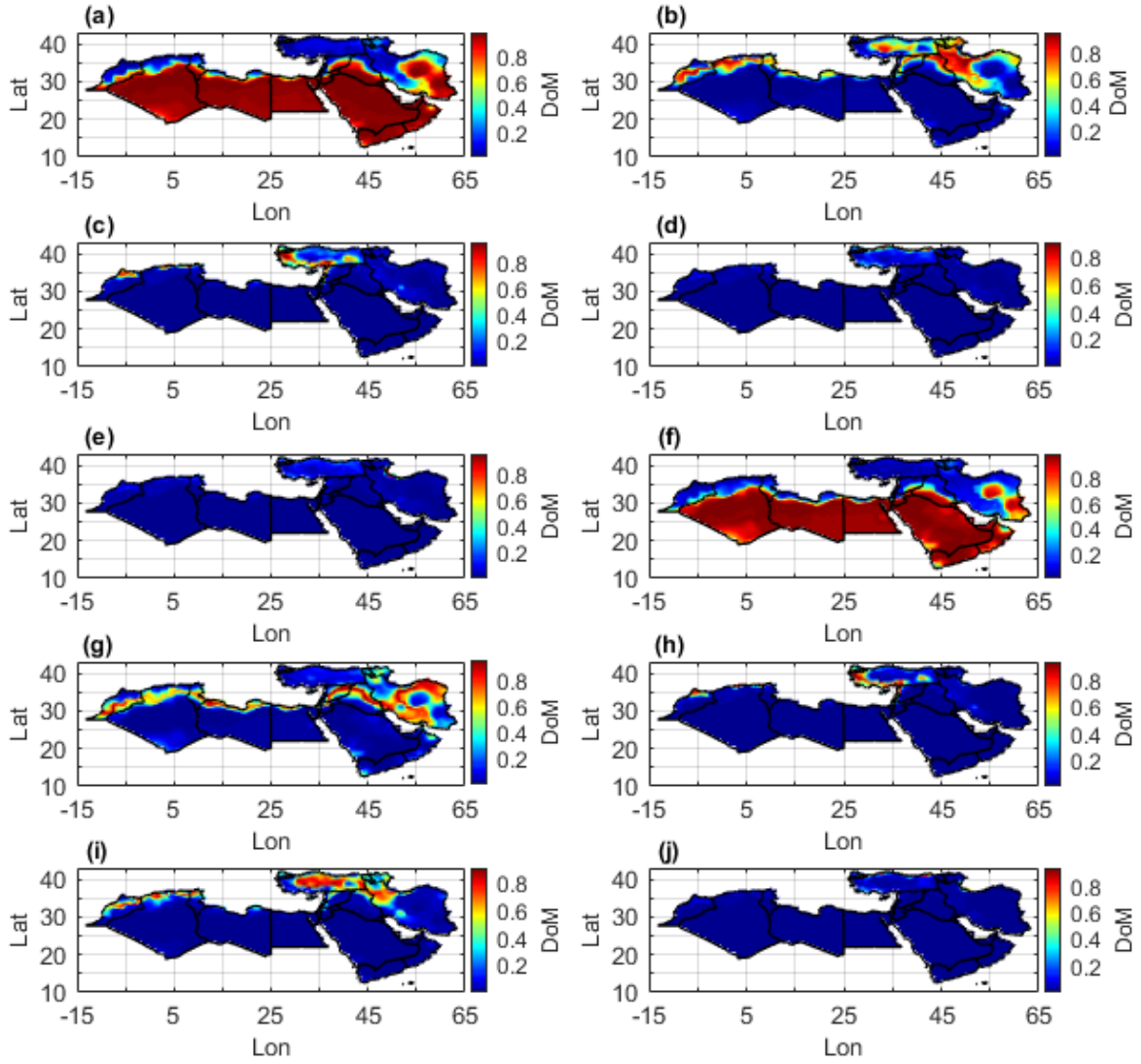

Fig. 15 Degree of Membership of the CPC clusters (daily, Cd and monthly, Cm) including (a) Cd1, (b) Cd2, (c) Cd3, (d) Cd4, (e) Cd5, (f) Cm1, (g) Cm2, (h) Cm3, (i) Cm4, and (j) Cm5

The Köppen-Geiger Classification (KGC) method, updated by Kottek et al. (2006), defines six main precipitation classes (desert, steppe, summer-dry, fully humid, winter-dry, and monsoonal), and the first four classes cover the MENA region. These KGC classes support the optimal number of clusters identified by DB_R, whereas other validity indices, such as the CS index, overestimate the optimal number of clusters. Moreover, the optimal number of clusters determined by the new validity index is larger than those found by the DB and Silhouette indices for all datasets. The differences in long-term monthly precipitation (Fig. 8) and cumulative distribution functions (Fig. 9) among clusters confirm that the clusters obtained by DB_R are distinct, whereas the DB and Silhouette indices tend to overlook these differences.

The spatial pattern of precipitation clustering over MENA shows that the dry cluster (the first cluster, e.g., Cr1, Gp1, Pr1, Ud1, Cd1, and Cm1) is similar across all datasets except PREC/L (Fig. 7). This dry cluster is consistent with the desert precipitation class of the updated KGC (Kottek et al., 2006), where four main deserts of the study domain—the Sahara in North Africa, the Arabian Desert, Dasht-e Kavir, and Dasht-e Lut—are located. Furthermore, the spatial distribution of cluster Cr2 aligns with the steppe class of the KGC more closely than the corresponding clusters in other datasets (Gp2, Pr2, Ud2, Cd2, and Cm2). Although clusters Cr3, Cr4, Gp3, Gp4, Pr3, Ud3, Cd3, Cm3, and Cm4 match the characteristics of the summer-dry class of the KGC, the geographic distribution of clusters from CRU, GPCC, and CPCM better agrees with the summer-dry class of the KGC. The spatial distribution of the clusters (Fig. 7) and seasonal precipitation changes of the cluster centroids (Fig. 8) show that most of the wet

clusters are located in Iran (northern, western, and southwestern areas), Turkey (coastal regions), Azerbaijan, Armenia, and the northern parts of African countries along the Mediterranean coast. These spatial patterns indicate that CRU and GPCC identify four main precipitation zones over Iran. The GPCC precipitation clusters over Iran closely follow the precipitation patterns defined in previous studies (Darand & Daneshvar, 2014; Javanmard et al., 2010). Precipitation patterns over Iran are closely linked to topography, where the Caspian Sea and the Zagros and Alborz mountain ranges influence precipitation distribution. In addition, the CRU and GPCC precipitation clusters are more consistent with the precipitation zones of Turkey identified by (Unal et al., 2003). Results showed that there is no dry month along the Black Sea coast in northern and northeastern Turkey (e.g., see clusters Cr6, Gp5, and Gp6). Furthermore, wet clusters with maximum rainfall in winter (see clusters Cr3, Gp3) appear along the Mediterranean coastal regions in southern and southwestern Turkey, and drier clusters in southeastern and eastern Turkey (e.g., clusters Cr4, Gp4) are the main clusters covering the principal precipitation zones of Turkey.

The Degree of Membership (DoM) indicator shows that most cluster members have a DoM greater than 0.5 relative to their cluster centroids, with the exception of the Cd4 and Cd5 clusters. These high DoM values confirm that adding clusters beyond the three optimal clusters suggested by the DB or Silhouette indices (Table 4) is necessary to adequately represent precipitation patterns over the MENA region. Among all datasets, the CPCM precipitation clusters are the most distinct (as confirmed by high DoM values), indicating strong similarity between cluster members and their centroids. In all datasets, the dry cluster (Cr1, Gp1, Pr1, Ud1, Cd1, and Cm1) members exhibit the greatest DoMs. As the number of clusters increases, intermediate clusters with smaller DoMs can emerge between the fully wet and dry clusters. For instance, although long-term monthly precipitation of cluster Cr4 (see Fig. 7a) follows the seasonal pattern of clusters Cr2 and Cr3 (e.g., dry months in Jun, Jul, Aug), the monthly precipitation values of this cluster distinguish it from Cr2 and Cr3 (see Fig. 8a). Specifically, precipitation in Cr4 falls between that of Cr2 and Cr3, which are considered relatively dry and wet clusters, respectively. Thus, this intermediate cluster has a smaller DoM relative to the wet or dry clusters. Similarly, cluster Gp4 can be considered an intermediate cluster between Gp2 and Gp3.

The use of Euclidean distance as the similarity/dissimilarity measure indicates that the obtained clusters exhibit noticeable differences in their behavior (e.g. in mean values or seasonal patterns, see Fig. 8). In addition, the differences among their eCDFs (Fig. 9) show that the variances and ranges of precipitation differ across clusters. (Fovell, 1997) also confirmed that the Euclidean distance criterion can capture differences in mean values, variance, and correlation between two time series. According to Eq. 1, Euclidean distance computes the similarity/dissimilarity of two time series at corresponding time points, which can be significantly influenced by extreme values and nonconstant variance (heteroscedasticity) of the time series (Shirkhorshidi et al., 2015). For instance, extreme precipitation values affect the distance between wet and dry clusters, causing the centroids of wet and relatively drier clusters to be closer. Thus, extremes that are more evident in the daily data (CPCD) compared to the monthly ones (CPCM) can affect intermediate clusters and lead to a wrong clustering of some data points. In addition, the lack of intermediate clusters in CPCD results in more homogeneous clustering [i.e., similar characteristics among clusters, such as in long-term monthly precipitation (Fig. 8e) or eCDFs (Fig. 9e and Fig. 14a)].

The developed cluster analysis can be effectively applied in many fields, including climate studies (e.g., for data assessment and preprocessing prior to analysis). Cluster analysis and the identification of climatic zones across different regions are valuable for mitigation and adaptation planning. Regions can be clustered based on the severity of climate impacts. Targeted management strategies can be applied to clusters facing extreme weather or hydrological events (e.g., heavy rainfall or drought), thereby reducing the risks associated with climate change.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Declaration of competing interests

All authors have no financial or personal relationships that could potentially bias the content or findings of this research.

Acknowledgments

The author would like to thank NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, *for making their datasets and documents publicly available. We also acknowledge the Iran Meteorological Organization (IRIMO) and the following modeling groups for making their datasets publicly available: the National Center for Atmospheric Research (NCAR), the Climatic Research Unit (CRU), the Global Precipitation Climatology Centre (GPCC), and the Climate Prediction Center (CPC).*

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern recognition*, *46*(1), 243-256. https://doi.org/https://doi.org/10.1016/j.patcog.2012.07.021

Bach, F. R., & Jordan, M. I. (2006). Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, *7*(Oct), 1963-2001. https://doi.org/https://doi.org/10.1002/9780470742044.ch13

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & Ziese, M. (2011). GPCC full data reanalysis version 6.0 at 0.5: monthly land-surface precipitation from rain-gauges built on GTS-based and historic data. *Global Precipitation Climatology Centre (GPCC): Berlin, Germany*. https://doi.org/https://doi.org/10.5676/dwd_gpcc/fd_m_v7_050

Chen, M., Xie, P., Janowiak, J. E., & Arkin, P. A. (2002). Global land precipitation: A 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, *3*(3), 249-266. https://doi.org/https://doi.org/10.1175/1525-7541(2002)003%3C0249:GLPAYM%3E2.0.CO;2

Chou, C.-H., Su, M.-C., & Lai, E. (2004). A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, *7*(2), 205-220. https://doi.org/https://doi.org/10.1007/s10044-004-0218-1

Darand, M., & Daneshvar, M. R. M. (2014). Regionalization of precipitation regimes in Iran using principal component analysis and hierarchical clustering analysis. *Environmental Processes*, *1*(4), 517-532. https://doi.org/https://doi.org/10.1007/s40710-014-0039-1

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224-227. https://doi.org/https://doi.org/10.1109/TPAMI.1979.4766909

de Vries, A. J., Tyrlis, E., Edry, D., Krichak, S., Steil, B., & Lelieveld, J. (2013). Extreme precipitation events in the Middle East: dynamics of the Active Red Sea Trough. *Journal of Geophysical Research: Atmospheres*, *118*(13), 7087-7108. https://doi.org/https://doi.org/10.1002/jgrd.50569

Fovell, R. G. (1997). Consensus clustering of US temperature and precipitation data. *Journal of Climate*, *10*(6), 1405-1427. https://doi.org/https://doi.org/10.1175/1520-0442(1997)010%3C1405:CCOUST%3E2.0.CO;2

Harris, I., & Jones, P. (2017). University of East Anglia Climatic Research Unit, CRU TS4. 01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901-Dec. 2016). *Centre for Environmental Data Analysis*, *4*. https://doi.org/http://dx.doi.org/10.5285/58a8802721c94c66ae45c3baa4d814d0

He, X., Zhang, S., & Liu, Y. (2015). An Adaptive Spectral Clustering Algorithm Based on the Importance of Shared Nearest Neighbors. *Algorithms*, *8*(2), 177-189. https://doi.org/https://doi.org/10.3390/a8020177

İnkaya, T., Kayalıgil, S., & Özdemirel, N. E. (2015). An adaptive neighbourhood construction algorithm based on density and connectivity. *Pattern Recognition Letters*, *52*, 17-24. https://doi.org/https://doi.org/10.1016/j.patrec.2014.09.007

Javanmard, S., Yatagai, A., Nodzu, M., BodaghJamali, J., & Kawamoto, H. (2010). Comparing high-resolution gridded precipitation data with satellite rainfall estimates of TRMM_3B42 over Iran. *Advances in Geosciences*, *25*, 119-125. https://doi.org/https://doi.org/10.5194/adgeo-25-119-2010

Kim, M., & Ramakrishna, R. (2005). New indices for cluster validity assessment. *Pattern Recognition Letters*, *26*(15), 2353-2363. https://doi.org/https://doi.org/10.1016/j.patrec.2005.04.007

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259-263.

https://doi.org/https://doi.org/10.1127/0941-2948/2006/0130

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). John Wiley & Sons.

Lhermitte, S., Verbesselt, J., Verstraeten, W. W., & Coppin, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote sensing of environment*, *115*(12), 3129-3152. https://doi.org/https://doi.org/10.1016/j.rse.2011.06.020

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874. https://doi.org/https://doi.org/10.1016/j.patcog.2005.01.025

McAfee, S., Guentchev, G., & Eischeid, J. (2014). Reconciling precipitation trends in Alaska: 2. Gridded data analyses. *Journal of Geophysical Research: Atmospheres*, *119*(24), 13,820-813,837. https://doi.org/https://doi.org/10.1002/2014JD022461

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems,

Nicholson, S. E., Some, B., McCollum, J., Nelkin, E., Klotter, D., Berte, Y., Diallo, B., Gaye, I., Kpabeba, G., & Ndiaye, O. (2003). Validation of TRMM and other rainfall estimates with a high-density gauge dataset for West Africa. Part I: Validation of GPCC rainfall product and pre-TRMM satellite and blended products. *Journal of Applied Meteorology*, *42*(10), 1337-1354. https://doi.org/https://doi.org/10.1175/1520-0450(2003)042%3C1337:VOTAOR%3E2.0.CO;2

Raziei, T., Bordi, I., & Pereira, L. S. (2011). An application of GPCC and NCEP/NCAR datasets for drought variability analysis in Iran. *Water resources management*, *25*(4), 1075-1086. https://doi.org/https://doi.org/10.1007/s11269-010-9657-1

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Schneider, U., Finger, P., Meyer-Christoffer, A., Rustemeier, E., Ziese, M., & Becker, A. (2017). Evaluating the hydrological cycle over land using the newly-corrected precipitation climatology from the Global Precipitation Climatology Centre (GPCC). *Atmosphere*, *8*(3), 52. https://doi.org/https://doi.org/10.3390/atmos8030052

Serra, J., & Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, *67*, 305-314. https://doi.org/http://dx.doi.org/10.1016/j.knosys.2014.04.035

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, *10*(12). https://doi.org/https://doi.org/10.1371/journal.pone.0144059

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79-107. https://doi.org/https://doi.org/10.1002/2017RG000574

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, *8*, 487-568.

Türkeş, M., & Tatlı, H. (2011). Use of the spectral clustering to determine coherent precipitation regions in Turkey for the period 1929–2007. *International Journal of Climatology*, *31*(14), 2055-2067. https://doi.org/https://doi.org/10.1002/joc.2212

Unal, Y., Kindap, T., & Karaca, M. (2003). Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *23*(9), 1045-1055. https://doi.org/https://doi.org/10.1002/joc.910

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416. https://doi.org/https://doi.org/10.1007/s11222-007-9033-z

Willmott, C. J. (2000). Terrestrial air temperature and precipitation: Monthly and annual time series (1950-1996). *http://climate.geog.udel.edu/~climate/html_pages/README.ghcn_ts.html*.

Xie, P., Chen, M., & Shi, W. (2010). CPC unified gauge-based analysis of global daily precipitation. Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc,

Yuan, C., Fan, K., & Sun, X. (2016). A self-adaptive spectral clustering based on geodesic distance and shared nearest neighbors. *Int J Hybrid Inf Technol*, *9*(4), 417-426.

https://doi.org/http://dx.doi.org/10.14257/ijhit.2016.9.4.36

Žalik, K. R., & Žalik, B. (2011). Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, *32*(2), 221-234. https://doi.org/https://doi.org/10.1016/j.patrec.2010.08.007

Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. Advances in neural information processing systems,

Zeng, S., Huang, R., Kang, Z., & Sang, N. (2014). Image segmentation using spectral clustering of Gaussian mixture models. *Neurocomputing*, *144*, 346-356. https://doi.org/https://doi.org/10.1016/j.neucom.2014.04.037

Zhang, X., Li, J., & Yu, H. (2011). Local density adaptive similarity measurement for spectral clustering. *Pattern Recognition Letters*, *32*(2), 352-358. https://doi.org/https://doi.org/10.1016/j.patrec.2010.09.014

Zhang, Z., Xu, C.-Y., El-Tahir, M. E.-H., Cao, J., & Singh, V. (2012). Spatial and temporal variation of precipitation in Sudan and their possible causes during 1948–2005. *Stochastic environmental research and risk assessment*, *26*(3), 429-441. https://doi.org/https://doi.org/10.1007/s00477-011-0512-6

Zhao, T., & Fu, C. (2006). Comparison of products from ERA-40, NCEP-2, and CRU with station data for summer precipitation over China. *Advances in Atmospheric Sciences*, *23*(4), 593-604. https://doi.org/https://doi.org/10.1007/s00376-006-0593-1