CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2617–2637
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Hydroclimatic Regionalization in Indonesia: A Comparative Study of Fuzzy C–Means and Gaussian Mixture Models

Bunga Anovia Putri Ramadhianti[1] and A'yunin Sofro*[2]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*
[2] *Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*

Abstract

Drought is a climate-related natural hazard that poses significant challenges to environmental sustainability and disaster risk reduction in Indonesia, particularly in the context of achieving Sustainable Development Goal 13 (Climate Action). However, drought-related characteristics vary considerably across regions due to differences in local climatic conditions. This study aims to regionalize drought-related hydroclimatic characteristics across 34 provinces in Indonesia using Fuzzy C-Means (FCM) and Gaussian Mixture Models (GMM). The analysis used climatological data from CHIRPS Daily and ERA5-Land covering the 2015–2025 period. Four indicators were considered, including rainfall, average temperature, number of rainy days, and consecutive dry days (CDD). The optimal number of clusters was determined using the Xie–Beni Index for FCM and the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for GMM, supported by cluster stability assessment based on the Adjusted Rand Index and additional validation using the Gap Statistic. Both methods identified two clusters as the optimal solution, consistently separating a relatively drier group of provinces from a relatively wetter group. Cluster quality was evaluated using the Silhouette Coefficient and Davies–Bouldin Index (DBI). The results showed that FCM provided marginally better clustering performance, with a Silhouette Coefficient of 0.345 and a DBI of 1.076, compared with 0.332 and 1.115, respectively, for GMM, although both methods produced broadly consistent regionalization. The resulting clusters reveal long-term drought-related regional patterns across Indonesia, providing a quantitative basis for identifying provinces with similar hydroclimatic characteristics to support further, more targeted drought-related research.

Keywords: Fuzzy C-Means; Gaussian Mixture Models; hydroclimatic characteristics; regionalization; SDG 13.

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

Drought is a climatic phenomenon that can adversely affect water availability, agricultural activities, ecosystems, and socioeconomic conditions. In Indonesia, drought-related conditions vary considerably across regions due to spatial and temporal differences in rainfall patterns and climatic factors. Indonesian rainfall variability is also influenced by large-scale climate phenomena such as the El Niño–Southern Oscillation (ENSO) and the Indian Ocean Dipole

*Corresponding author. E-mail: ayuninsofro@unesa.ac.id

Received: September 1, 2026 Revised: October 21, 2026 Accepted: November 2, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.45896

(IOD) [1]. These variations indicate that drought-related conditions across Indonesia cannot be adequately represented as a uniform condition. Therefore, regionalization based on similarities in hydroclimatic characteristics is needed to provide a more differentiated understanding of drought-related conditions.

Regionalization groups areas according to similarities in selected characteristics and can provide contextual information relevant to drought risk assessment and context-specific water, climate, and agricultural management. This regionalization reflects similarities in hydroclimatic characteristics rather than geographic contiguity. Thus, provinces belonging to the same cluster may not be geographically adjacent. A multivariate approach allows several hydroclimatic characteristics to be considered simultaneously, providing a broader representation of drought-related conditions than a single indicator. This approach is also relevant to the Sustainable Development Goals (SDGs), particularly SDG 6 on sustainable water management and SDG 13 on climate action [2], [3]. Previous studies have demonstrated the application of multivariate regionalization to identify areas with relatively similar drought and hydroclimatic characteristics [4].

Several clustering approaches have been applied to hydroclimatic and environmental regionalization. Fuzzy C-Means (FCM) assigns a degree of membership to each observation for every cluster, allowing regional characteristics to be represented through varying degrees of cluster association [5], [6]. In contrast, the Gaussian Mixture Model (GMM) represents observations as a mixture of Gaussian distributions and estimates the probability of each observation belonging to each component [7], [8], [9]. These approaches have been applied in hydrological regionalization, ecohydrological zoning, and environmental classification [5], [7], [10]. In Indonesia, clustering and regionalization have been used to classify provinces according to drought risk based on daily maximum temperature [11], classify rainfall patterns using time-series characteristics [12], and regionalize dry-season rainfall for drought risk management in East Java [13]. These studies demonstrate the applicability of clustering methods to hydroclimatic analysis, but differ in terms of spatial coverage, observation periods, and the characteristics used to construct regional groups.

Drought-related hydroclimatic conditions are not solely associated with low rainfall but can also be reflected in temperature, the frequency of rainy days, and the duration of consecutive dry periods. Consecutive Dry Days (CDD) is defined as the duration of consecutive days with precipitation below a specified threshold [14]. Although these indicators provide complementary information on drought-related conditions, regionalization across all provinces of Indonesia using multiple indicators within a comparative fuzzy and probabilistic clustering framework remains limited. In particular, the different mechanisms of FCM and GMM provide an opportunity to examine whether fuzzy membership-based and probability-based clustering approaches produce different regional structures when applied to the same multivariate hydroclimatic characteristics.

Therefore, this study aims to classify the 34 provinces of Indonesia according to drought-related hydroclimatic characteristics, represented by their long-term average conditions over the 2015–2025 period, using FCM and GMM. This focus emphasizes persistent hydroclimatic patterns rather than interannual variability or extreme drought events. Four indicators are considered: rainfall, mean temperature, number of rainy days, and CDD. The optimal number of clusters for FCM is determined using the Xie–Beni Index, while AIC and BIC are used to determine the optimal number of GMM components. The resulting cluster structures are evaluated using the Silhouette Coefficient and Davies–Bouldin Index to quantitatively compare the two approaches. The contribution of this study lies in the comparative regionalization of all Indonesian provinces using multiple drought-related hydroclimatic indicators under fuzzy and probabilistic clustering frameworks, providing a quantitative basis for identifying regional similarities and assessing the suitability of the two approaches for drought-related hydroclimatic regionalization.

2. Methods

This section describes the data sources and preprocessing procedures, the Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM) clustering methods, and the use of the gap statistic to support the selection of the optimal number of clusters. It also presents the criteria for evaluating clustering quality and stability, as well as the overall research workflow. The procedures are described in the sequence used in the analysis.

2.1. Data and Preprocessing

2.1.1. Data Source

This study utilized climatological data from the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) Daily dataset [15] and the ERA5-Land Daily Aggregated dataset [16], covering the period from 2015 to 2025. The CHIRPS dataset ($\approx$ 5.5 km spatial resolution) was used to derive rainfall-related indicators, including total rainfall, the number of rainy days, and Consecutive Dry Days (CDD), while ERA5-Land ($\approx$ 11 km spatial resolution) was used to obtain mean air temperature. The unit of analysis consisted of the 34 provinces of Indonesia, based on the geoBoundaries ADM1 administrative boundary dataset [17]. This boundary version reflects the 34-province structure predating Indonesia's most recent provincial reorganization (e.g., the further division of Papua into additional provinces), and was applied consistently across all four indicators. Grid-based CHIRPS and ERA5-Land values were aggregated to the provincial level by averaging all grid cells within each provincial polygon (zonal mean) using the `reduceRegions` function in Google Earth Engine. No missing values were encountered in the resulting provincial-level dataset for any of the four indicators. Data extraction and initial processing were conducted using Google Earth Engine, while subsequent data preprocessing, statistical analysis, and clustering procedures were performed using Google Colab.

Four climatological indicators were calculated annually for each province throughout the study period. The annual values were then averaged to obtain representative values for the 2015–2025 period, which were used as input variables for the clustering. This temporal averaging was adopted to characterize the long-term hydroclimatic baseline of each province rather than to capture interannual drought dynamics. Consequently, this approach does not reflect extreme drought years or interannual variability. Rainfall was defined as the accumulated precipitation on days classified as rainy (RR $\geq$ 1 mm/day), consistent with the commonly used rain-day threshold in climatological studies [18]. This definition differs from total precipitation, which would also include trace amounts below the 1 mm threshold; the rainy-day-based accumulation was adopted for consistency with the number-of-rainy-days and CDD indicators, which apply the same threshold. The number of rainy days was determined by counting days with RR $\geq$ 1 mm, while Consecutive Dry Days (CDD) were defined as the maximum number of consecutive days with RR $<$ 1 mm within each year, averaged across the study period at the provincial level. Mean temperature was calculated as the average daily air temperature (temperature_2m) converted from Kelvin to degrees Celsius. Thus, the final dataset consisted of 34 provincial observations and four variables, namely rainfall, average temperature, the number of rainy days, and consecutive dry days (CDD), as presented in Table 1.

Table 1: Average values of drought-related indicators across Indonesian provinces during 2015–2025.

No.	Province	Rainfall (mm)	Average Temperature (°C)	Number of Rainy Days (day)	CDD (day)
1	Aceh	2733.653	22.568	243.117	10.586
2	Bali	2149.561	24.221	123.125	31.589
3	Bangka–Belitung Islands	3015.498	26.799	203.882	19.185
4	Banten	3296.532	25.589	206.965	20.993
⋮	⋮	⋮	⋮	⋮	⋮
34	West Sumatra	3054.664	23.571	226.880	12.684

2.1.2. Z–score Standardization

Before the clustering process, all research variables are standardized using Z-score standardization to place variables with different units and ranges on a comparable scale. Standardization is commonly applied in cluster analysis because differences in measurement scales may cause variables with larger numerical ranges to have a disproportionate influence on the clustering process [19], [20]. The Z-score standardization is defined in Eq. (1).

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{1}$$

where z_{ij} denotes the standardized value of the i-th observation for the j-th variable, x_{ij} is the original value of the i-th observation for the j-th variable, $\bar{x}_j$ is the mean of the j-th variable, and s_j is its standard deviation. This transformation produces standardized variables with a mean of zero and a standard deviation of one, allowing each variable to contribute more comparably to the clustering process.

2.2. Fuzzy C–Means

2.2.1. Fuzzy C–Means Formulation

Fuzzy C–Means (FCM) is a clustering method based on fuzzy set theory that assigns each observation a degree of membership in every cluster. Unlike hard clustering, which assigns an observation exclusively to a single cluster, FCM allows an observation to belong simultaneously to multiple clusters with different membership degrees [21], [22]. Let the dataset be represented by $X = \{x_1, x_2, \ldots, x_n\}$, where n denotes the number of observations. The membership degrees are represented by a partition matrix $U = [u_{ik}]$, where u_{ik} denotes the membership degree of observation i in cluster k. The membership values satisfy Eq. (2):

$$0 \leq u_{ik} \leq 1, \quad i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, K \tag{2}$$

with the normalization constraint Eq. (3):

$$\sum_{k=1}^{K} u_{ik} = 1, \quad i = 1, 2, \ldots, n \tag{3}$$

where K is the number of clusters [22], [6]. The FCM algorithm determines the partition matrix and cluster centers by minimizing the objective function in Eq. (4).

$$J_m(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m \left\| x_i - v_k \right\|^2 \tag{4}$$

where $J_m(U, V)$ is the objective function, n is the number of observations, K is the number of clusters, u_{ik} is the membership degree of observation i in cluster k, $m > 1$ is the fuzziness parameter, x_i denotes the i-th observation, and v_k is the center of cluster k. The objective function represents the weighted within-cluster variation, where the contribution of each observation is weighted according to its membership degree [22], [6].

Given the membership matrix, the center of the k-th cluster is updated according to Eq. (5).

$$v_k = \frac{\sum_{i=1}^{n} u_{ik}^m x_i}{\sum_{i=1}^{n} u_{ik}^m} \tag{5}$$

Subsequently, the membership degree of each observation is recalculated according to its relative distance to the cluster centers using Eq. (6).

$$u_{ik} = \frac{1}{\sum_{l=1}^{K} \left(\frac{\|x_i - v_k\|}{\|x_i - v_l\|} \right)^{\frac{2}{m-1}}} \tag{6}$$

where l indexes the clusters used in the normalization of the membership degrees. Thus, an observation located closer to a particular cluster center generally receives a higher membership degree in that cluster. The updates of the cluster centers and membership degrees are performed alternately until a convergence criterion is satisfied. Convergence can be determined based on the change in the membership matrix or the objective function between consecutive iterations falling below a predefined tolerance value, or when the maximum number of iterations is reached [22], [6]. Through this iterative optimization process, FCM produces a soft partition in which an observation may exhibit different degrees of association with multiple clusters. For subsequent interpretation, each observation may be assigned to the cluster corresponding to its highest membership degree.

In this study, the fuzziness parameter was set to $m = 2$, which is the most commonly used value in FCM applications [22]. The algorithm was run with a convergence tolerance of 1×10^{-5} and a maximum of 1000 iterations. FCM was applied for candidate cluster numbers ranging from $K = 2$ to $K = 10$. For each candidate K, FCM was independently run with 50 different random seeds (0 – 49) to assess the sensitivity of the resulting partition to random initialization. The final FCM model was fitted using a fixed random seed of 0 to ensure reproducibility.

2.2.2. Xie–Beni Index

The Xie–Beni Index (XB) is used to determine the optimal number of clusters in the Fuzzy C–Means method. This index evaluates the quality of fuzzy clustering by considering two main characteristics: cluster compactness and separation between clusters. The XB value is obtained by comparing the total variation of observations relative to their cluster centers with the minimum distance between cluster centers [23]. Mathematically, the XB index is defined as follows:

$$XB = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K} u_{ij}^{m} \left\| x_i - v_j \right\|^2}{n \min_{j \neq k} \left\| v_j - v_k \right\|^2} \tag{7}$$

where n denotes the number of observations, K represents the number of clusters, u_{ij} is the membership degree of the i-th observation in the j-th cluster, m is the fuzziness parameter, x_i denotes the i-th observation, v_j represents the center of the j-th cluster, and v_k represents the center of the k-th cluster. A smaller XB value indicates a clustering result with greater within-cluster compactness and better separation between clusters. Therefore, the optimal number of clusters is selected based on the minimum XB value among the alternative numbers of clusters evaluated [23].

2.2.3. Cluster Stability Assessment

To ensure that the selected number of clusters does not merely reflect a particular random initialization, cluster stability was assessed using multiple independent runs of the FCM algorithm for $K = 2$ to 10. For each candidate number of clusters, 50 independent initializations were generated using different random seeds, each producing a distinct partition of the observations. The consistency between partitions was measured using the Adjusted Rand Index (ARI), which quantifies the agreement between two partitions while correcting for the level of agreement expected by chance [24], [25]. For each candidate K, the mean pairwise ARI across the 50 resulting partitions was computed as a stability score, where a higher mean ARI indicates a more consistent (stable) partition across initializations. The number of clusters for FCM was selected primarily based on the minimum Xie–Beni (XB) value calculated using Eq. (7), while the mean pairwise ARI was used as complementary evidence to assess the robustness of the resulting partition to random initialization. This approach allows the selected solution to be evaluated not only in terms of cluster compactness and separation but also in terms of stability across different initializations.

2.3. Gaussian Mixture Models

2.3.1. Gaussian Mixture Models Formulation

The Gaussian Mixture Model (GMM) is a probabilistic clustering method that represents the distribution of observations as a finite mixture of Gaussian components. Each component corresponds to a potential cluster and is characterized by a mixing proportion, mean vector, and covariance matrix. Unlike hard clustering methods, GMM describes the association of an observation with each component through posterior probabilities, allowing an observation to have non-zero probabilities of belonging to multiple clusters [8], [9]. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a dataset consisting of n observations. A GMM with K components models the probability density of an observation x_i as defined in Eq. (8).

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{8}$$

where K denotes the number of Gaussian components or clusters, π_k is the mixing proportion of the k-th component, and $\mathcal{N}\left(x_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ denotes a multivariate Gaussian probability density function with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The mixing proportions satisfy Eq. (9):

$$\sum_{k=1}^{K} \pi_k = 1, \qquad \pi_k \geq 0. \tag{9}$$

Thus, the overall distribution is represented as a weighted combination of K Gaussian distributions [8], [9].

The parameters of a GMM, consisting of π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, are commonly estimated using the iterative Expectation-Maximization (EM) algorithm. In the Expectation step, the posterior probability, commonly referred to as the responsibility, that observation x_i belongs to component k is calculated using Eq. (10).

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}\left(x_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{l=1}^{K} \pi_l \mathcal{N}\left(x_i \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} \tag{10}$$

Here, γ_{ik} denotes the posterior probability that the i-th observation belongs to the k-th component. For each observation, the posterior probabilities satisfy Eq. (11):

$$\sum_{k=1}^{K} \gamma_{ik} = 1. \tag{11}$$

During the Maximization step, the model parameters are updated based on the posterior probabilities obtained in the Expectation step. First, the effective number of observations associated with component k is calculated using Eq. (12):

$$N_k = \sum_{i=1}^{n} \gamma_{ik}. \tag{12}$$

The updated mixing proportion, mean vector, and covariance matrix are then given by Eq. (13), Eq. (14), and Eq. (15), respectively:

$$\pi_k = \frac{N_k}{n}, \tag{13}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} x_i, \tag{14}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} \left(x_i - \boldsymbol{\mu}_k\right) \left(x_i - \boldsymbol{\mu}_k\right)^T. \tag{15}$$

The Expectation and Maximization steps are performed alternately until convergence is achieved. The convergence of the EM algorithm can be assessed through the change in the log-likelihood between consecutive iterations or when a predefined maximum number of iterations is reached [8], [9]. After convergence, each observation may be assigned to the component with the highest posterior probability for subsequent cluster interpretation.

In this study, four covariance structures were considered for $\mathbf{\Sigma}_k$: full, diagonal, tied, and spherical, differing in how much flexibility is allowed in modeling the shape and orientation of each component. A regularization term of 1×10^{-6} was used to improve numerical stability. For each $K = 2$ to $K = 10$ and covariance structure, the EM algorithm used `n_init`=1, `init_params`=`kmeans`, a tolerance of 1×10^{-3}, and a maximum of 1000 iterations. For each combination of K and covariance structure, 50 independent runs were performed using random seeds 1000–1049. The final GMM model was fitted using a fixed random seed of 0 to ensure reproducibility.

2.3.2. Akaike Information Criterion

The Akaike Information Criterion (AIC) is used as one of the criteria for determining the optimal number of components in a GMM. AIC considers the trade-off between model fit and model complexity. Thus, adding more components does not necessarily result in a better model when the improvement in model fit is not sufficient to justify the additional complexity [26]. The AIC is defined in Eq. (16):

$$\mathrm{AIC} = -2\ln(L) + 2p \tag{16}$$

where L denotes the maximum likelihood of the model and p represents the number of estimated parameters. According to this criterion, the model with the smaller AIC value is considered more appropriate.

2.3.3. Bayesian Information Criterion

In addition to AIC, the Bayesian Information Criterion (BIC) is used to evaluate the number of components in a GMM. Similar to AIC, BIC considers both model fit and model complexity, but it imposes a larger penalty for additional parameters, particularly as the number of observations increases [27]. The BIC is defined in Eq. (17):

$$\mathrm{BIC} = -2\ln(L) + p\ln(n) \tag{17}$$

where L denotes the maximum likelihood of the model, p represents the number of estimated parameters, and n denotes the number of observations. The optimal number of components is determined by selecting the model with the lowest BIC value.

2.3.4. Cluster Stability Assessment

In addition to AIC and BIC, cluster stability was assessed to evaluate the robustness of the GMM solutions to different random seeds. For each candidate number of components ($K = 2$ to $K = 10$), 50 independent runs were performed using the diagonal covariance structure, with `n_init`=1 and random seeds ranging from 1000 to 1049. The GMM parameters were initialized using the K-means initialization strategy. The EM algorithm used a convergence tolerance of 10^{-3}, a maximum of 1000 iterations, and a covariance regularization parameter of 10^{-6}. The consistency of the resulting partitions was assessed using the Adjusted Rand Index (ARI), which quantifies the agreement between two partitions while correcting for the level of agreement expected by chance [24], [25]. Each run was compared with the first run as the reference solution, and the mean ARI across the remaining runs was used as a measure of assignment stability, with higher values indicating greater consistency in cluster assignments.

Because GMM also provides probabilistic cluster assignments, the stability of the estimated diagonal variance parameters and posterior probabilities was additionally assessed. The relative difference between matched diagonal variance parameters was used to quantify the consistency of

the estimated component variances across runs, with smaller values indicating greater parameter stability. Posterior probability stability was evaluated using the mean absolute error (MAE) between aligned posterior probability estimates across runs, where smaller values indicate more consistent probabilistic assignments. The convergence rate across independent runs was also recorded to assess whether the fitted models consistently reached convergence. The stability assessment was used as complementary evidence for evaluating the robustness of candidate solutions, while the selection of the number of components was primarily based on model-selection criteria, particularly AIC and BIC, together with cluster validity and interpretability.

2.4. Gap Statistic

The Gap Statistic proposed by Tibshirani et al. [28] was additionally employed to validate the selected number of clusters. The method compares the within-cluster dispersion of the observed data with that expected under reference datasets. In this study, 500 reference datasets were generated using the scaled principal component approach. The optimal number of clusters was determined using the one-standard-error criterion, which selects the smallest K satisfying Eq. (18):

$$\mathrm{Gap}(K) \geq \mathrm{Gap}(K+1) - SE(K+1). \tag{18}$$

The Gap Statistic was applied as an additional validation criterion for both FCM and GMM clustering. Its use is consistent with recent climate and hydroclimatic regionalization studies that have employed the Gap Statistic alongside other clustering validity measures to determine the number of clusters [29].

2.5. Cluster Evaluation

In this study, clustering quality was evaluated using the Silhouette Coefficient and the Davies–Bouldin Index. Because both indices are defined based on hard cluster assignments, the Silhouette Coefficient and Davies–Bouldin Index are calculated using the hard partition obtained from FCM and GMM, in which each observation is assigned to the cluster with the highest membership degree (for FCM) or posterior probability (for GMM). The resulting hard assignments are used for both evaluation indices.

2.5.1. Silhouette Coefficient

The Silhouette Coefficient is used to evaluate clustering quality by comparing the similarity of an observation to other observations within its assigned cluster with its similarity to observations in neighboring clusters. For each observation, the silhouette value is determined based on the average intra-cluster distance and the average distance to the nearest neighboring cluster [30]. Let C_r denote the cluster containing observation x_i, and let $|C_r|$ represent the number of observations in that cluster. The average intra-cluster distance of x_i, denoted by $a(i)$, is defined in Eq. (19):

$$a(i) = \frac{1}{|C_r| - 1} \sum_{\substack{x_j \in C_r \\ j \neq i}} d(x_i, x_j) \tag{19}$$

where $d(x_i, x_j)$ denotes the distance between observations x_i and x_j. Thus, $a(i)$ measures the average dissimilarity between observation x_i and all other observations belonging to the same cluster. The value $b(i)$ is then defined as the minimum of these average distances across all clusters other than C_r, as shown in Eq. (20):

$$b(i) = \min_{s \neq r} \left\{ \frac{1}{|C_s|} \sum_{x_j \in C_s} d(x_i, x_j) \right\} \tag{20}$$

Thus, $b(i)$ represents the average distance between observation x_i and the nearest neighboring cluster. Based on $a(i)$ and $b(i)$, the silhouette value for the i-th observation is calculated using

Eq. (21).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{21}$$

The silhouette value ranges from -1 to 1. A value close to 1 indicates that an observation is well matched to its assigned cluster and clearly separated from other clusters. A value close to 0 suggests overlap between clusters, whereas a negative value indicates that an observation may be closer to another cluster than to the cluster to which it has been assigned. The overall Silhouette Coefficient is obtained by averaging the silhouette values across all n observations, as expressed in Eq. (22).

$$SC = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{22}$$

A higher Silhouette Coefficient indicates better clustering quality [30].

2.5.2. Davies–Bouldin Index

The Davies–Bouldin Index (DBI) is used to evaluate clustering quality by considering both the compactness of individual clusters and the degree of separation between clusters. The index compares the dispersion of observations within a cluster with the distance between the centroids of different clusters [31]. Let C_i denote the i-th cluster, c_i its centroid, and n_i the number of observations in the cluster. The within-cluster dispersion, denoted by S_i, is defined as the average distance between observations in cluster C_i and its centroid, as shown in Eq. (23):

$$S_i = \frac{1}{n_i}\sum_{x \in C_i} d(x, c_i) \tag{23}$$

where $d(x, c_i)$ denotes the distance between observation x and the centroid c_i. A smaller value of S_i indicates a more compact cluster. The separation between clusters C_i and C_j is measured by the distance between their centroids, as shown in Eq. (24):

$$M_{ij} = d(c_i, c_j) \tag{24}$$

Based on the within-cluster dispersions and the separation between clusters, the similarity measure between clusters C_i and C_j is defined in Eq. (25):

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{25}$$

For each cluster C_i, the maximum similarity with all other clusters is then determined. The Davies–Bouldin Index is calculated as the average of these maximum similarity values across all K clusters, as expressed in Eq. (26).

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left\{\frac{S_i + S_j}{M_{ij}}\right\} \tag{26}$$

where K denotes the number of clusters. A lower DBI value indicates better clustering performance, reflecting more compact clusters and greater separation between clusters [31].

2.6. Research Workflow

The research workflow consists of sequential stages, beginning with the extraction of climatological data from the CHIRPS Daily and ERA5-Land datasets. Four drought-related hydroclimatic indicators, namely rainfall, mean temperature, the number of rainy days, and Consecutive Dry Days (CDD), are constructed for the 2015–2025 period and aggregated to obtain representative values for each of the 34 provinces. The resulting variables are standardized using Z-score standardization before clustering.

Two clustering approaches are then applied independently: Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM). For each method, multiple candidate numbers of clusters (or components) are evaluated using method-specific criteria, namely the Xie–Beni Index for FCM and AIC and BIC for GMM, in combination with a cluster stability assessment based on the Adjusted Rand Index (ARI) across 50 independent random initializations. The optimal number of clusters is determined based on these criteria and further supported by the Gap Statistic using the one-standard-error criterion. The selected cluster solutions are subsequently evaluated using the Silhouette Coefficient and Davies–Bouldin Index. Finally, the resulting regionalization patterns and cluster characteristics are compared between FCM and GMM to examine their representation of drought-related hydroclimatic characteristics across Indonesia.

3. Results and Discussion

This section presents the hydroclimatic characteristics of the 34 Indonesian provinces and the determination of the optimal number of clusters. It then discusses the regionalization results obtained using Fuzzy C-Means (FCM) and Gaussian Mixture Models (GMM), followed by an evaluation of clustering quality, stability, and spatial patterns. The comparisons use the indicators and evaluation measures described in the Methods section.

3.1. Descriptive Analysis of Hydroclimatic Characteristics

Descriptive statistical analysis was conducted to provide an initial overview of the drought-related characteristics across the 34 provinces in Indonesia. Table 2 summarizes the four indicators. These statistics provide context for the clustering results discussed below.

Table 2: Descriptive statistics of drought-related indicators.

Statistic	Rainfall (mm)	Average Temperature (°C)	Rainy Days (Day)	CDD (Day)
Mean	2784.982	24.741	191.222	21.197
Std	499.994	1.206	35.805	8.086
Min	1723.870	21.979	104.457	9.785
25%	2504.304	23.897	170.592	15.030
50%	2835.825	24.890	199.403	19.288
75%	3049.688	25.528	217.727	25.042
Max	3720.357	27.130	249.343	45.001

As presented in Table 2, rainfall exhibited considerable variation across provinces, ranging from 1,723.870 to 3,720.357 mm, with an average of 2,784.982 mm. The number of rainy days ranged from 104.457 to 249.343 days, with a mean of 191.222 days. In contrast, average temperature showed relatively lower variability, ranging from 21.979 to 27.130 °C, with an average of 24.741 °C. Meanwhile, consecutive dry days (CDD) ranged from 9.785 to 45.001 days, with a mean of 21.197 days. These differences in the ranges and distributions of the indicators indicate substantial variation in climatic conditions and drought-related characteristics among provinces in Indonesia.

The boxplots presented in Figure 1 provide a further overview of the distribution of the drought-related indicators. Based on the interquartile range (IQR) criterion, rainfall, average temperature, and the number of rainy days showed no potential outliers. In contrast, the CDD variable contained one potential outlier, with a value of approximately 45 days, corresponding to East Nusa Tenggara Province. This value indicates that East Nusa Tenggara experienced a longer period of consecutive dry days than most other provinces. Such differences highlight the heterogeneity of wet and dry conditions across Indonesian provinces, providing an important basis for the subsequent clustering analysis.

The correlation analysis in Figure 2 shows a strong negative relationship between rainy days and CDD (-0.87), and rainfall correlates strongly with rainy days (0.71) and negatively with CDD (-0.64), indicating that wetter provinces with more rainy days tend to have shorter dry

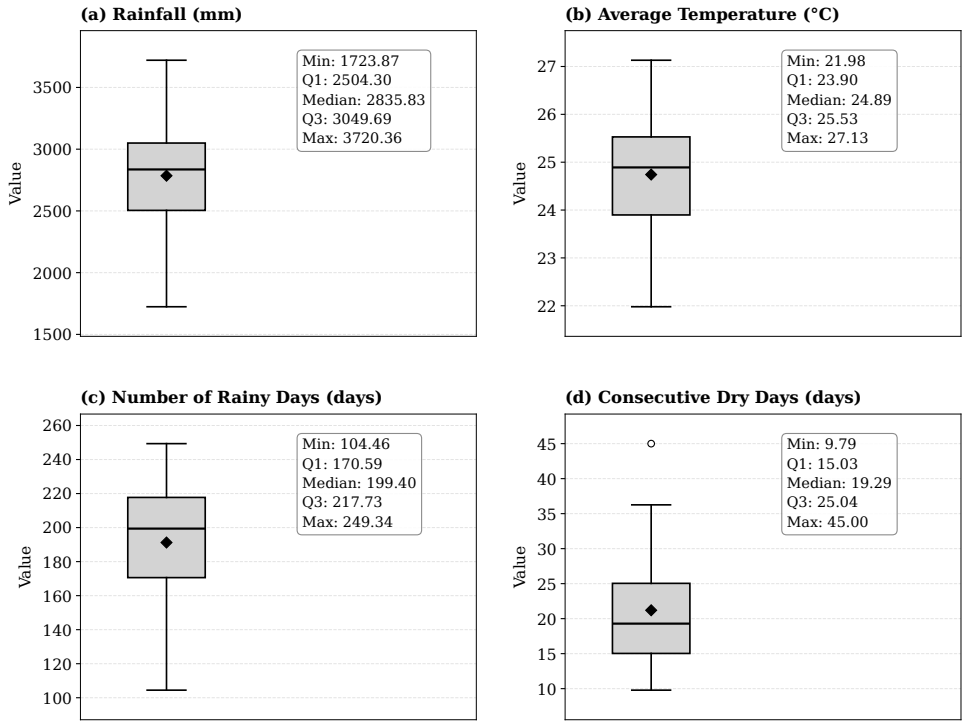


Fig. 1: Boxplots of the four hydroclimatic indicators across the 34 provinces in Indonesia.

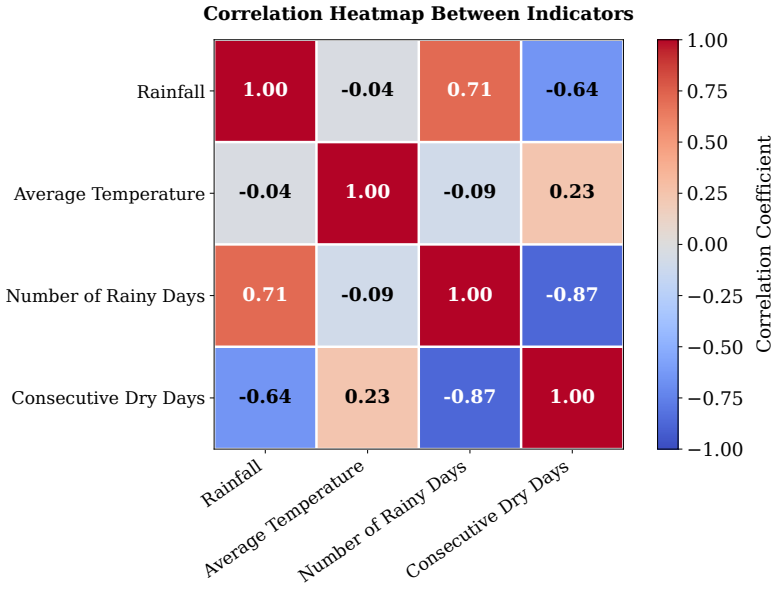


Fig. 2: Correlation heatmap between drought-related indicators.

spells. Average temperature, however, shows only weak linear relationships with the other three variables, suggesting it captures a distinct climatic pattern.

To assess whether these strong correlations imply redundancy in clustering, a variable sensitivity analysis was performed by individually removing each variable from the baseline FCM configuration (all four variables, $K = 2$) and comparing the resulting partitions to the baseline using ARI, Silhouette, DBI, and XB (Table 3). Removing rainfall, rainy days, or CDD each substantially reduced ARI (0.572, 0.667, and 0.771, respectively), showing that despite their strong correlations, each variable still contributes non-redundant information for province grouping. Removing average temperature left the cluster structure unchanged (ARI = 1.000) while slightly improving Silhouette, DBI, and XB values, indicating its inclusion has minimal effect

on grouping but slightly reduces cluster separation. All four variables were therefore retained in the final configuration to preserve the multidimensional characterization of drought-related hydroclimatic conditions.

Table 3: Variable sensitivity analysis using FCM ($K = 2$).

Scenario	No. of Variables	ARI vs Baseline	Silhouette	DBI	XB
Baseline (4 variables)	4	1.000	0.346	1.076	0.265
Without Rainfall	3	0.572	0.341	1.088	0.292
Without Temperature	3	1.000	0.453	0.814	0.144
Without Rainy Days	3	0.667	0.335	1.066	0.331
Without CDD	3	0.772	0.362	1.081	0.279

3.2. Determination of the Optimal Number of Clusters

The optimal number of clusters for the Fuzzy C–Means (FCM) method was evaluated based primarily on the Xie–Beni Index (XB), with the mean pairwise Adjusted Rand Index (ARI) used as complementary evidence of partition stability across candidate cluster numbers ranging from $K = 2$ to $K = 10$. For each candidate K, 50 independent random initializations were evaluated. These runs also provided the partitions used to assess stability.

Table 4: Results of the Xie–Beni index and ARI stability score for FCM.

K	XB (mean)	ARI (mean)
2	0.2646	1.000
3	0.3311	1.000
4	0.2647	1.000
5	0.2692	0.988
6	0.3349	0.812
7	0.3218	0.826
8	0.5639	0.728
9	0.6694	0.739
10	0.6521	0.674

As shown in Table 4, $K = 2$, $K = 3$, and $K = 4$ each achieved a mean ARI of 1.000, indicating perfectly stable partitions across the 50 random initializations. However, the stability decreased for larger values of K, particularly from $K = 6$ onward. Among the perfectly stable solutions, the lowest Xie–Beni Index was obtained at $K = 2$ (XB = 0.2646), followed very closely by $K = 4$ (XB = 0.2647). The near-identical XB values indicate that both solutions exhibit comparable compactness and separation, while $K = 2$ provides the lower XB value with equally stable cluster assignments. As an additional validation, the Gap Statistic was applied using the one-standard-error criterion, which selected $K = 2$ with $\text{Gap}(2) = 0.262$ and $SE(2) = 0.090$. Therefore, considering the Xie–Beni Index, clustering stability, and the additional Gap Statistic validation, $K = 2$ was selected as the final FCM solution.

For the Gaussian Mixture Model (GMM), the optimal number of components was evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and mean Adjusted Rand Index (ARI) stability score across candidate values from $K = 2$ to $K = 10$, comparing four covariance structures (full, diagonal, tied, and spherical) with 50 independent random initializations per candidate K. The full covariance structure showed AIC and BIC decreasing monotonically without a clear minimum, along with consistently lower ARI, indicating overfitting due to the large number of estimated parameters relative to the limited sample size (34 provinces) and less reproducible cluster assignments. The diagonal covariance structure, in contrast, produced a clear BIC minimum and higher ARI, reflecting more stable and reproducible clustering. Based on this comparison, the diagonal covariance structure was adopted for the final model.

Table 5: AIC, BIC, and ARI stability results for GMM.

K	AIC (mean)	BIC (mean)	ARI (mean)
2	368.718	394.664	0.8549
3	364.392	404.077	0.5982
4	353.248	406.671	0.6831
5	354.284	421.444	0.6200
6	346.565	427.462	0.6133
7	342.119	436.753	0.6197
8	334.095	442.467	0.6646
9	323.679	445.788	0.6836
10	306.271	442.117	0.6846

As presented in Table 5, the lowest BIC value was obtained at $K = 2$ (BIC = 394.664), which also exhibited the highest mean pairwise ARI among the candidate solutions (ARI = 0.8549). The combination of the lowest BIC and the highest assignment stability provides support for the two-component solution as a parsimonious and robust representation of the data. Although $K = 10$ yielded the lowest AIC (306.271), its BIC was substantially higher than that of $K = 2$ (442.117 versus 394.664), while its mean ARI was lower (0.6846 versus 0.8549). The continuous decrease in AIC with increasing K indicates that AIC favors increasingly complex models for these data, whereas BIC imposes a stronger penalty for model complexity and identifies $K = 2$ as the preferred solution. As an additional validation, the Gap Statistic also selected $K = 2$ using the one-standard-error criterion, with $\text{Gap}(2) = 0.235$ and $SE(2) = 0.113$. This result provides additional support for the two-component GMM solution identified based on the information criteria.

3.3. Regionalization Using Fuzzy C-Means

The regionalization results obtained using the Fuzzy C-Means (FCM) method with the optimal number of two clusters are presented in Table 6. The table reports membership values for each cluster and the resulting assignment. These values provide the basis for describing the provincial groups below.

Table 6: Membership values and cluster assignment.

No.	Province	Membership_C1	Membership_C2	Cluster_FCM
1	Aceh	0.7511	0.2488	1
2	Bali	0.1094	0.8905	2
3	Bangka–Belitung Islands	0.6860	0.3139	1
4	Banten	0.8557	0.1442	1
⋮	⋮	⋮	⋮	⋮
34	West Sumatra	0.8926	0.1073	1

Unlike hard clustering methods, FCM assigns each province a degree of membership in every cluster. Each province is subsequently assigned to the cluster for which it has the highest membership value. The results indicate that the cluster sizes are unequal, with 21 provinces assigned to Cluster 1 and 13 provinces assigned to Cluster 2. To further describe the differences in climatic and drought characteristics among the clusters, the average values of the four variables within each cluster are presented in Table 7.

Table 7: Number of provinces and average characteristics of each cluster obtained using FCM.

Cluster	Number of Provinces	Average Rainfall (mm)	Average Temperature (°C)	Average Number of Rainy Days (Day)	Average Consecutive Dry Days (Day)
1	21	3069.661	24.617	213.769	16.807
2	13	2325.117	24.941	154.800	28.288

The cluster profiles reveal notable differences in climatic and drought characteristics between the two groups, consistent with the moderate degree of cluster separation indicated by the Silhouette Coefficient obtained for this configuration. Cluster 1, consisting of 21 provinces, is characterized by relatively wetter conditions, with a higher average rainfall of 3069.661 mm and a greater number of rainy days, averaging 213.769 days. This cluster also records the lowest average CDD value of 16.807 days, indicating shorter consecutive dry periods. In contrast, Cluster 2, consisting of 13 provinces, exhibits relatively drier characteristics, with a lower average rainfall of 2325.117 mm and fewer rainy days, averaging 154.800 days. This cluster also has a higher average CDD value of 28.288 days, indicating longer consecutive dry periods, along with a slightly higher average temperature of 24.941 °C compared with Cluster 1. The composition of provinces within each cluster is presented in Table 8, and the corresponding spatial distribution is illustrated in Figure 3. The map uses the WGS 84 geographic coordinate reference system (EPSG:4326), with provincial boundaries following the 34-province structure described in Section 2.1.1.

Table 8: List of provinces from the FCM clustering method.

Cluster	Number of Provinces	Province
1	21	East Kalimantan, Jambi, West Papua, West Sumatra, Papua, Bengkulu, Central Kalimantan, West Kalimantan, Banten, North Sumatra, North Kalimantan, South Sumatra, South Kalimantan, Lampung, Riau, Aceh, West Java, Bangka-Belitung Islands, South Sulawesi, Central Java, West Sulawesi
2	13	East Java, Special Region of Yogyakarta, Bali, North Sulawesi, West Nusa Tenggara, Gorontalo, Southeast Sulawesi, East Nusa Tenggara, Jakarta Special Capital Region, Riau Islands, North Maluku, Maluku, Central Sulawesi

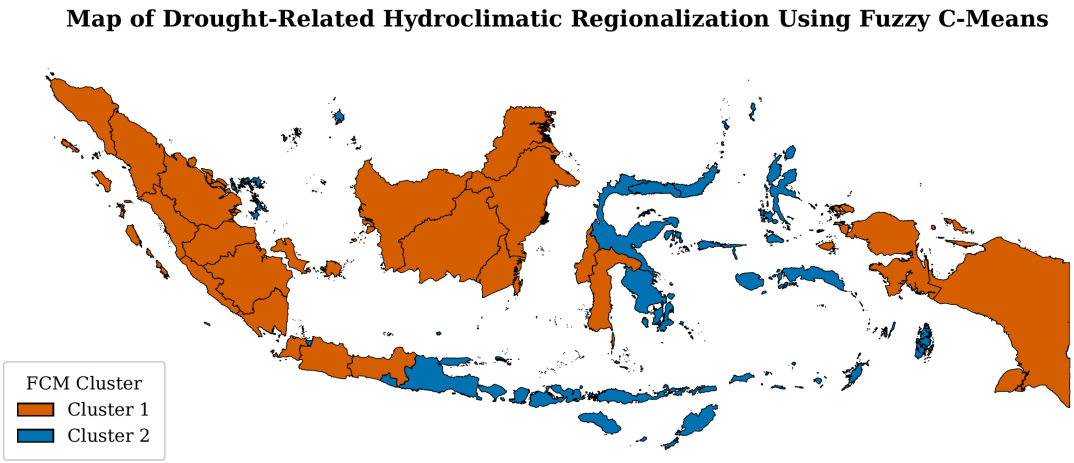


Fig. 3: Map of drought-related hydroclimatic regionalization using FCM.

Cluster 1 comprises the majority of provinces in Sumatra, Kalimantan, and the Papua region, along with Banten, West Java, Central Java, South Sulawesi, and West Sulawesi, consistent with its relatively wetter profile. Cluster 2 includes Bali, West Nusa Tenggara, and East Nusa Tenggara, consistent with the relatively drier characteristics of this cluster, along with several provinces in Sulawesi and Maluku, as well as Jakarta and East Java, indicating that the drier profile is not confined to a single geographic region. As shown in Figure 3, provinces assigned to the same cluster are not necessarily geographically contiguous, indicating that the regionalization is primarily driven by similarities in climatic and drought characteristics rather than geographical proximity. These results indicate that variations in drought characteristics among Indonesian provinces cannot be described by a single climatic indicator but rather reflect the combined influence of rainfall, average temperature, the number of rainy days, and CDD.

3.4. Regionalization Using Gaussian Mixture Model

The regionalization results obtained using the Gaussian Mixture Model (GMM) with the optimal number of two components are presented in Table 9. The table reports component probabilities and the resulting cluster assignment. These outputs provide the basis for the cluster summaries below.

Table 9: Gaussian Mixture Model clustering results.

No.	Province	Probability_C1	Probability_C2	Cluster_GMM
1	Aceh	0.000747	9.99E-01	2
2	Bali	1.000	9.89E-12	1
3	Bangka–Belitung Islands	0.011215	9.89E-01	2
4	Banten	0.004929	9.95E-01	2
⋮	⋮	⋮	⋮	⋮
34	West Sumatra	0.000327	9.99E-01	2

Unlike conventional hard clustering approaches, GMM estimates the probability of each province belonging to every mixture component. Each province is subsequently assigned to the cluster with the highest membership probability. The clustering results show an unequal distribution of provinces across the two clusters, with 15 provinces assigned to Cluster 1 and 19 provinces assigned to Cluster 2.

The average characteristics of each cluster are presented in Table 10 to describe the differences in climatic and drought conditions between the two groups. The table includes cluster size and the mean value of each indicator. These summaries are used to compare the two cluster profiles.

Table 10: Number of provinces and average characteristics of each cluster obtained using GMM.

Cluster	Number of Provinces	Average Rainfall (mm)	Average Temperature (°C)	Average Number of Rainy Days (Day)	Average Consecutive Dry Days (Day)
1	15	2384.360	24.745	159.052	27.476
2	19	3101.263	24.738	216.619	16.239

The results indicate differences in climatic and drought characteristics between the two clusters, consistent with the moderate degree of cluster separation reflected in the corresponding Silhouette Coefficient. Cluster 1, consisting of 15 provinces, is characterized by relatively drier conditions, with a lower average rainfall of 2384.360 mm and fewer rainy days, averaging 159.052 days. This cluster also records a higher average CDD of 27.476 days, indicating longer consecutive dry periods. In contrast, Cluster 2, consisting of 19 provinces, exhibits relatively wetter characteristics, with a higher average rainfall of 3101.263 mm and a greater number of rainy days, averaging 216.619 days, along with a lower average CDD of 16.239 days. The average temperature is nearly identical between the two clusters (24.745°C for Cluster 1 versus 24.738°C for Cluster 2), indicating that the separation between clusters is driven primarily by rainfall-related characteristics rather than temperature.

To assess the robustness of the two-component GMM solution to different random initializations, a stability analysis was conducted. All 50 runs converged successfully, yielding a convergence rate of 100%. Cluster assignment stability was evaluated using the pairwise Adjusted Rand Index (ARI) across all 1,225 pairs of runs. The mean ARI was 0.8549 (SD = 0.1378), with values ranging from 0.6688 to 1.0000, indicating that the province assignments were generally consistent across different initializations, although some variation in cluster membership remained.

The stability of the diagonal variance parameters was also assessed using their mean relative difference across pairwise runs. The mean relative difference was 0.0589 (SD = 0.0567), with a maximum of 0.1452, indicating relatively small variation in the estimated variance parameters across different initializations. Furthermore, the stability of the posterior probabilities was

evaluated using the mean absolute error (MAE) between aligned posterior probabilities across runs. The mean posterior MAE was 0.0237 (SD = 0.0235), with a maximum value of 0.0563, indicating that the estimated posterior probabilities were generally consistent across different initializations. Overall, these findings indicate that the two-component GMM solution is reasonably robust to changes in random initialization, supporting the reliability of the resulting regionalization. The composition of provinces within each cluster is presented in Table 11, and the corresponding spatial distribution is illustrated in Figure 4. The map uses the WGS 84 geographic coordinate reference system (EPSG:4326), with provincial boundaries following the 34-province structure described in Section 2.1.1.

Table 11: List of provinces from the GMM clustering method.

Cluster	Number of Provinces	Province
1	15	East Nusa Tenggara, West Nusa Tenggara, Bali, East Java, North Sulawesi, Gorontalo, Riau Islands, Special Region of Yogyakarta, Jakarta Special Capital Region, Maluku, North Maluku, Southeast Sulawesi, Central Sulawesi, West Sulawesi, Central Java
2	19	West Kalimantan, Papua, Central Kalimantan, North Kalimantan, West Sumatra, West Papua, North Sumatra, Bengkulu, Aceh, Riau, Jambi, South Sumatra, Banten, East Kalimantan, Bangka-Belitung Islands, South Kalimantan, Lampung, West Java, South Sulawesi

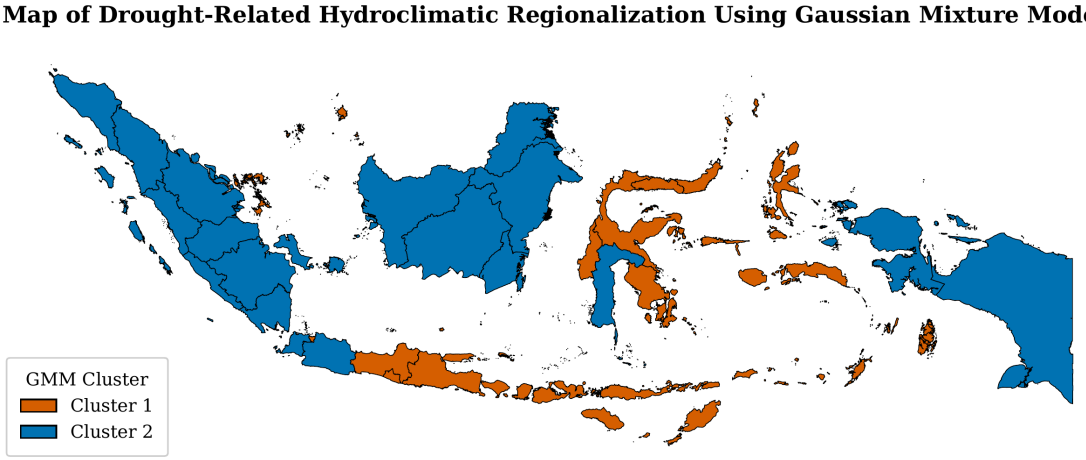


Fig. 4: Map of drought-related hydroclimatic regionalization using GMM.

Cluster 1 includes Bali, West Nusa Tenggara, and East Nusa Tenggara, consistent with its relatively drier profile, together with several provinces in Sulawesi and Maluku, as well as Jakarta, East Java, and Central Java, indicating that the drier profile spans several non-contiguous regions of Indonesia. Cluster 2 comprises the majority of provinces in Sumatra and Kalimantan, along with Papua, West Papua, West Java, and South Sulawesi, consistent with its relatively wetter profile. As shown in Figure 4, provinces assigned to the same cluster are not necessarily geographically contiguous, indicating that the grouping is primarily determined by similarities in climatic and drought characteristics rather than geographical proximity. This spatial pattern highlights the ability of GMM to identify provinces with similar multivariate climatic characteristics despite geographical differences. These results are broadly consistent with the FCM regionalization, with both methods separating a relatively drier group centered on Bali and Nusa Tenggara from a relatively wetter group comprising most of Sumatra and Kalimantan.

3.5. Evaluation and Comparison of Clustering Results

The quality of the clustering results obtained using the Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM) methods was evaluated using the Silhouette Coefficient and the Davies–Bouldin Index (DBI). Table 12 presents both measures for the two methods. Higher Silhouette values and lower DBI values indicate more favorable cluster structure.

Table 12: Comparison of clustering evaluation results between FCM and GMM.

Method	Silhouette Coefficient	Davies–Bouldin Index
Fuzzy C–Means (FCM)	0.345	1.076
Gaussian Mixture Model (GMM)	0.332	1.115

The evaluation results presented in Table 12 indicate that FCM achieved a higher Silhouette Coefficient of 0.345 compared with 0.332 obtained by GMM. A higher Silhouette Coefficient indicates that observations are, on average, more closely associated with their assigned clusters than with neighboring clusters. In addition, FCM produced a lower DBI value of 1.076, whereas the GMM result was 1.115. A lower DBI indicates a more favorable clustering structure, characterized by greater within-cluster compactness and clearer separation between clusters. Although the difference between the two methods is relatively small, FCM produced marginally better clustering quality than GMM for the data analyzed in this study according to both evaluation criteria.

Both methods identified two clusters as the optimal configuration, consistently separating a relatively drier group from a relatively wetter group. However, the resulting cluster sizes and composition differed between the two methods. FCM produced a 21–13 split (Cluster 1: relatively wetter, Cluster 2: relatively drier), whereas GMM produced a 19–15 split (Cluster 2: relatively wetter, Cluster 1: relatively drier). This difference indicates that, while both methods agree on the broad wet–dry distinction, they classify a number of provinces near the boundary between the two profiles differently. To quantify the agreement between the final FCM and GMM partitions, the Adjusted Rand Index (ARI) was calculated. The resulting ARI of 0.7719 indicates substantial agreement between the two regionalization results, despite differences in the assignment of some provinces. This result supports the consistency of the broad wet–dry regionalization identified by both methods.

Despite these differences, both methods showed a generally consistent pattern in identifying provinces with relatively dry and wet characteristics. For example, Bali, West Nusa Tenggara, and East Nusa Tenggara were consistently classified into the relatively drier cluster by both FCM and GMM. Conversely, most provinces in Sumatra and Kalimantan, which are characterized by higher rainfall and a greater number of rainy days together with lower CDD values, were consistently grouped into the relatively wetter cluster by both methods.

Table 13: Comparison of cluster characteristics obtained using FCM and GMM.

Characteristics	**FCM**	**GMM**
Relatively drier	Cluster 2 (13 provinces): lower rainfall and number of rainy days, higher CDD	Cluster 1 (15 provinces): lower rainfall and number of rainy days, higher CDD
Relatively wetter	Cluster 1 (21 provinces): higher rainfall and number of rainy days, lower CDD	Cluster 2 (19 provinces): higher rainfall and number of rainy days, lower CDD

Table 13 summarizes the correspondence between the cluster characteristics identified by the two methods. Both FCM and GMM consistently distinguished a relatively drier group from a relatively wetter group based on the same set of climatic and drought characteristics. The main difference between the two methods lies in the exact composition of provinces assigned to each group, with FCM assigning a somewhat larger share of provinces to the wetter cluster (21 out of 34) compared with GMM (19 out of 34). Combined with its marginally higher Silhouette Coefficient and lower DBI, FCM can be regarded as providing a slightly more compact and better-separated regionalization for the data and indicators considered in this study, although

both methods yield a broadly consistent picture of drought-related hydroclimatic conditions across Indonesian provinces.

4. Conclusion

This study examined drought-related hydroclimatic characteristics across 34 Indonesian provinces based on long-term (2015–2025) averages of rainfall, mean temperature, rainy days, and Consecutive Dry Days (CDD), using Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM). Based on method-specific validity indices and a cluster stability assessment using the Adjusted Rand Index, both methods identified two clusters as optimal, consistently separating a relatively drier group of provinces from a relatively wetter one. FCM showed marginally better cluster quality than GMM (Silhouette Coefficient: 0.345 vs. 0.332; Davies–Bouldin Index: 1.076 vs. 1.115), though both methods produced broadly consistent results. A variable sensitivity analysis further showed that, despite strong correlations among rainfall, rainy days, and CDD, each variable contributed non-redundant information to the cluster structure, supporting the retention of all four indicators.

These findings provide a quantitative basis for identifying provinces with similar long-term hydroclimatic characteristics. However, several limitations should be noted: the clusters reflect long-term tendencies rather than validated drought risk classes, given the use of 11-year averages and the absence of external drought validation, the analysis is based on a relatively small sample of 34 provinces under a specific administrative boundary configuration that does not reflect Indonesia's most recent provincial reorganization, and geographic proximity was not incorporated into the clustering, so resulting groups need not be spatially contiguous. Future research may address these limitations by incorporating drought-specific indices such as SPI or SPEI, accounting for spatial dependence, and validating the resulting regionalization against independent drought impact data.

CRediT Authorship Contribution Statement

Bunga Anovia Putri Ramadhianti: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, Writing Original Draft Preparation. **A'yunin Sofro:** Methodology, Validation, Supervision, Writing–Review & Editing.

Declaration of Generative AI and AI-assisted technologies

Generative AI was used during the preparation of this manuscript to assist with language editing, translation, and organization of the manuscript. ChatGPT (GPT-5.6 Luna) was used as a writing assistance tool. The authors reviewed and verified all AI-assisted content and remain fully responsible for the accuracy, interpretation, and originality of the manuscript. No generative AI tool was used to perform the statistical analysis or generate the research findings.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

This study did not involve human participants, human data, or animal subjects. All hydroclimatic data used in this study (CHIRPS Daily and ERA5-Land) are publicly available secondary datasets. Ethics approval was not required for this study.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The datasets used in this study were obtained from the publicly available CHIRPS Daily and ERA5-Land datasets. The materials required to reproduce the analysis, including the processed provincial-level data, data extraction and preprocessing procedures, and scripts used for the Fuzzy C–Means (FCM) and Gaussian Mixture Model (GMM) analyses, are available from the corresponding author upon reasonable request. These materials include the parameter settings, random seeds, initialization strategies, convergence tolerance, maximum iterations, and covariance specifications used in the analyses. The code is not publicly available but can be obtained from the corresponding author upon reasonable request.

References

[1] R. Hidayat and K. Ando. "Variabilitas Curah Hujan Indonesia dan Hubungannya dengan ENSO/IOD: Estimasi Menggunakan Data JRA-25/JCDAS". In: *Agromet* 28.1 (2014), pp. 1–8. DOI: 10.29244/j.agromet.28.1.1-8.

[2] United Nations Department of Economic and Social Affairs. *Water and Sanitation*. Online. Accessed: Sep. 11, 2026. 2026. https://sdgs.un.org/topics/water-and-sanitation.

[3] United Nations Department of Economic and Social Affairs. *Climate Action and Synergies*. Online. Accessed: Sep. 11, 2026. 2026. https://sdgs.un.org/topics/climate-action-synergies.

[4] M. Azam, H. K. Park, S. J. Maeng, and H. S. Kim. "Regionalization of Drought across South Korea Using Multivariate Methods". In: *Water* 10.1 (2018), p. 24. DOI: 10.3390/w10010024.

[5] S. Sadri and D. H. Burn. "A Fuzzy C-Means approach for regionalization using a bivariate homogeneity and discordancy approach". In: *Journal of Hydrology* 401.3–4 (2011), pp. 231–239. DOI: 10.1016/j.jhydrol.2011.02.027.

[6] J. C. Bezdek, R. Ehrlich, and W. Full. "FCM: The fuzzy c-means clustering algorithm". In: *Computers and Geosciences* 10.2–3 (1984), pp. 191–203. DOI: 10.1016/0098-3004(84)90020-7.

[7] M. Bai, W. Li, and F. Li. "Ecohydrological zoning based on the Gaussian mixture model clustering method". In: *Journal of Hydroinformatics* 27.4 (2025), pp. 771–786. DOI: 10.2166/hydro.2025.305.

[8] G. J. McLachlan and S. U. Chang. "Mixture modelling for cluster analysis". In: *Statistical Methods in Medical Research* 13.5 (2004), pp. 347–361. DOI: 10.1191/0962280204sm372ra.

[9] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. "Finite mixture models". In: *Annual Review of Statistics and Its Application* 6 (2019), pp. 355–378. DOI: 10.1146/annurev-statistics-031017-100325.

[10] M. Kempf. "Global to regional ecozone classifications from multivariate clustering models". In: *Ecological Informatics* 96 (2026), p. 103840. DOI: 10.1016/j.ecoinf.2026.103840.

[11] A. Sofro, E. R. Auliya, K. N. Khikmah, D. Ariyanto, R. W. Romadhonia, and H. Ombao. "Clustering Provinces with Drought Risk Based on Daily Maximum Temperature". In: *International Journal of Environmental Science and Development* 15.2 (2024), pp. 69–76. DOI: 10.18178/ijesd.2024.15.2.1470.

[12] A. Sofro, R. A. Riani, K. N. Khikmah, R. W. Romadhonia, and D. Ariyanto. "Analysis of Rainfall in Indonesia Using a Time Series-Based Clustering Approach". In: *Barekeng* 18.2 (2024), pp. 0837–0848. DOI: 10.30598/barekengvol18iss2pp0837-0848.

[13] H. Mulyanti, I. Istadi, and R. Gernowo. "Comprehensive regionalization of dry season rainfall for enhanced drought risk management in monsoonal region of Indonesia". In: *Stochastic Environmental Research and Risk Assessment* 40.2 (2026). DOI: 10.1007/s00477-026-03169-3.

[14] Y. Duan, Z. Ma, and Q. Yang. "Characteristics of consecutive dry days variations in China". In: *Theoretical and Applied Climatology* 130.1 (2017), pp. 701–709. DOI: 10.1007/s00704-016-1984-6.

[15] C. Funk et al. "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes". In: *Scientific Data* 2.1 (2015), p. 150066. DOI: 10.1038/sdata.2015.66.

[16] J. Muñoz-Sabater et al. "ERA5-Land: a state-of-the-art global reanalysis dataset for land applications". In: *Earth System Science Data* 13.9 (2021), pp. 4349–4383. DOI: 10.5194/essd-13-4349-2021.

[17] D. Runfola et al. "GeoBoundaries: A global database of political administrative boundaries". In: *PLoS ONE* 15.4 (2020). DOI: 10.1371/journal.pone.0231866.

[18] Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). *Laju Perubahan Curah Hujan - Iklim - BMKG*. Online. Accessed: Sep. 11, 2026. https://www.bmkg.go.id/iklim/analisis-laju-perubahan-curah-hujan.

[19] C. M. Schaffer and P. E. Green. "An Empirical Comparison of Variable Standardization Methods in Cluster Analysis". In: *Multivariate Behavioral Research* 31.2 (1996), pp. 149–167. DOI: 10.1207/s15327906mbr3102_1.

[20] G. W. Milligan and M. C. Cooper. "A study of standardization of variables in cluster analysis". In: *Journal of Classification* 5.2 (1988), pp. 181–204. DOI: 10.1007/bf01897163.

[21] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046.

[22] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, 1981. DOI: 10.1007/978-1-4757-0450-1.

[23] X. L. Xie and G. Beni. "A validity measure for fuzzy clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.8 (1991), pp. 841–847. DOI: 10.1109/34.85677.

[24] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. DOI: 10.1007/bf01908075.

[25] D. Steinley. "Properties of the Hubert-Arabie adjusted Rand index". In: *Psychological Methods* 9.3 (2004), pp. 386–396. DOI: 10.1037/1082-989x.9.3.386.

[26] H. Akaike. "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/tac.1974.1100705.

[27] G. Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: 10.1214/aos/1176344136.

[28] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the Number of Clusters in a Data Set Via the Gap Statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423. DOI: 10.1111/1467-9868.00293.

[29] M. H. Rahman and H. Sadia. "Prognosticate the Analogous Region in Bangladesh Utilizing an Unsupervised Machine Learning Technique". In: *International Journal of Data Science and Analysis* 11.3 (2025), pp. 46–62. DOI: 10.11648/j.ijdsa.20251103.11.

[30] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[31] D. L. Davies and D. W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: 10.1109/tpami.1979.4766909.