



INTERPRETABLE UNSUPERVISED IDENTIFICATION OF PERSISTENT RAINFALL REGIMES WITH ROBUSTNESS AND TEMPORAL STABILITY ASSESSMEN

Imane Meguenni*¹ and Bendaoud Mebarek²

^{1, 2}LRIAS – Artificial Intelligence and Systems Research Laboratory University of Tiaret, Tiaret, Algeria.

¹<https://orcid.org/0009-0007-5546-9155>^{ID}, ²<https://orcid.org/0000-0002-6838-3867>^{ID}

Email: *imane.meguenni@univ-tiaret.dz, mebarekbendaoud@yahoo.fr

ARTICLE INFO

Article History

Received: March 7, 2026

Reviewed: June 12, 2026

Accepted: July 12, 2026

Published: July 31, 2026

Keywords:

Rainfall regimes,
Unsupervised learning,
Climate regionalization,
Robust clustering,
Temporal stability.

ABSTRACT

Rainfall variability strongly constrains agricultural planning, drought-risk management, and climate resilience. Yet many large-scale rainfall regime studies provide limited evidence on robustness across clustering methods and on temporal persistence of the discovered regimes. This study proposes an interpretable unsupervised artificial intelligence framework to identify dominant rainfall regimes across 36 Indian meteorological subdivisions using monthly rainfall records spanning 1901–2015. We engineered transparent climatic descriptors capturing rainfall magnitude, interannual variability, seasonal peak intensity, and intra-annual concentration. K-Means clustering with silhouette-based model selection favored a two-regime partition ($k = 2$), achieving a Silhouette coefficient of 0.428, indicating moderate but meaningful separability. Robustness was evaluated by applying Ward hierarchical clustering to the same standardized feature space, yielding comparable separation (Silhouette = 0.413) and strong cross-method agreement (Adjusted Rand Index, ARI = 0.881). Statistical separability was confirmed using one-way ANOVA with large effect sizes (η^2 up to 0.860 for mean annual rainfall), highlighting annual magnitude and peak intensity as primary discriminative drivers. Temporal replication across three historical sub-periods (1901–1950, 1951–1980, 1981–2015) showed high persistence: 35 of 36 subdivisions retained stable regime membership, with only Lakshadweep transitioning once ($P1 \rightarrow P2$). The resulting pipeline is reproducible, physically interpretable, and directly relevant for climate-risk zoning and long-term water-resource and agricultural planning.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Rainfall variability is a fundamental driver of agricultural productivity, hydrological balance, and climate-risk exposure in monsoon-dominated regions such as India. Long-term precipitation dynamics directly influence crop calendars, irrigation demand, groundwater recharge, and drought vulnerability. Recent assessments by the Intergovernmental Panel on Climate Change (IPCC) highlight increasing variability in precipitation patterns and a rising frequency of hydro-climatic extremes, particularly in vulnerable regions [1], [2]. Evaluating structural persistence is crucial to differentiate persistent hydro-climatic organization from transient variability, especially in light of the documented long-term climate variability over India [3]. These developments reinforce the need for robust rainfall regionalization frameworks capable of identifying persistent structural patterns rather than short-term fluctuations. Climate regionalization aims to delineate spatially homogeneous areas sharing similar atmospheric or hydro-climatic characteristics [4]. Within this conceptual framework, a rainfall regime can be defined as a spatially coherent pattern characterized by comparable rainfall magnitude, interannual variability, and seasonal concentration over an extended temporal horizon. Reliable identification of such regimes is essential for agricultural zoning, water-resource allocation, and long-term climate adaptation planning. India provides a particularly suitable case study due to its strong monsoonal system and pronounced regional rainfall contrasts.

High-resolution rainfall datasets across Indian meteorological subdivisions [5], together with globally consistent reanalysis products such as ERA5 [6], offer the opportunity to investigate long-term precipitation organization over more than a century. These datasets allow objective examination of rainfall magnitude, variability, and seasonality across 1901–2015 within a unified analytical framework. Traditional rainfall classifications have frequently relied on expert-defined regional boundaries or synoptic-scale atmospheric typologies [4]. While climatologically informative, such approaches may overlook latent multivariate structures embedded in long-term precipitation records. In contrast, unsupervised learning techniques enable data-driven discovery of intrinsic spatial organization without predefined labels, making them particularly appropriate for large hydro-climatic datasets [7]. Among unsupervised clustering methods, K-Means remains widely adopted due to its computational efficiency, conceptual simplicity, and suitability for moderate-dimensional climate descriptors [8]. The algorithm partitions observations by minimizing within-cluster variance relative to cluster centroids, thereby uncovering latent multivariate structure. However, clustering results may be sensitive to initialization and the selected number of clusters. Rigorous validation procedures are therefore necessary to ensure that identified rainfall regimes reflect meaningful hydro-climatic organization rather than methodological artifacts. Internal validation metrics such as the Silhouette coefficient quantify the degree of cohesion and separation between clusters [9]. Additionally, agreement between alternative clustering solutions can be evaluated using the Adjusted Rand Index (ARI), which measures similarity between partitions while correcting for chance agreement [10]. Together, these metrics provide complementary perspectives on structural robustness.

Beyond algorithmic validation, statistical inference provides further scientific rigor. One-way Analysis of Variance (ANOVA) tests whether rainfall descriptors differ significantly across identified clusters, while effect size measures such as eta squared (η^2) quantify the proportion of total variance explained by regime membership [11]. Large η^2 values indicate strong structural separation rather than merely statistically significant but practically negligible differences. Integrating clustering, robustness validation, and effect-size assessment strengthens interpretability and reduces the risk of overinterpreting arbitrary partitions. A further unresolved issue in rainfall regionalization concerns temporal stability. Many studies focus on spatial classification at a single time scale without systematically examining whether identified regimes persist across historical sub-periods. Given documented long-term climate variability over India [3], assessing structural persistence becomes essential to distinguish enduring hydro-climatic organization from transient variability. Despite numerous rainfall clustering studies, relatively few investigations simultaneously integrate: (i) interpretable feature engineering grounded in physically meaningful rainfall descriptors, (ii) cross-method robustness validation using ARI, (iii) statistical effect-size quantification, and (iv) explicit multi-period temporal stability assessment within a unified long-term framework. Accordingly, this study addresses the following research question:

Can long-term Indian rainfall (1901–2015) be robustly structured into statistically separable and temporally persistent hydro-climatic regimes using an interpretable unsupervised learning framework?

The objectives of this work are to:

- 1.Engineer physically interpretable rainfall descriptors capturing magnitude, variability, seasonality, and concentration patterns.
- 2.Identify dominant rainfall regimes using silhouette-optimized K-Means clustering.
- 3.Evaluate cross-method robustness via Ward hierarchical clustering and ARI.
- 4.Quantify statistical separability using ANOVA and effect-size metrics (η^2).
- 5.Assess regime persistence across three historical sub-periods.

By combining unsupervised discovery, robustness validation, statistical inference, and temporal replication, this study proposes a reproducible and physically interpretable rainfall regionalization framework suitable for climate-risk screening and long-term agricultural planning.

Organization of the Paper

The remainder of this paper is structured as follows. Section II presents the theoretical reference framework, including climate regionalization principles, unsupervised learning for climate Classification, clustering validation metrics, statistical effect-size measures, and concepts of temporal regime stability. Section III describes the dataset, feature engineering strategy, clustering methodology, robustness assessment procedures, and statistical validation approach in a fully reproducible manner. Section IV reports empirical results, including optimal cluster selection, regime interpretation, robustness analysis, statistical separability, temporal stability assessment and discusses the scientific implications, practical relevance for climate-risk zoning, and methodological limitations. Finally, Section V concludes the study and outlines potential directions for future research.

II. THEORETICAL REFERENCE

II.1 CLIMATE REGIONALIZATION AND RAINFALL REGIMES

Climate regionalization aims to identify spatially coherent areas sharing similar atmospheric or hydro-climatic characteristics [4]. In precipitation studies, a rainfall regime represents a spatially structured pattern defined by consistent rainfall magnitude, variability, and seasonal concentration over a long-term period [12]. Such classifications are particularly relevant in monsoon-dominated systems, where intra-annual rainfall concentration strongly influences agricultural productivity and hydrological planning [13]. Traditional rainfall regionalization approaches rely on expert-based classifications or predefined geographical partitions. However, these methods may not capture latent structures inherent in multivariate climate data. Data-driven approaches offer a more objective alternative by identifying homogeneous regions directly from observed climatic descriptors [14].

II.2 UNSUPERVISED LEARNING FOR CLIMATE CLASSIFICATION

Unsupervised learning techniques identify intrinsic patterns in datasets without labeled outputs. In climate science, clustering algorithms are widely used to regionalize rainfall, temperature, and atmospheric circulation indices [4], [15]. Among these methods, K-Means clustering remains one of the most applied techniques due to its computational efficiency and conceptual simplicity [16]. K-Means partitions a dataset into k clusters by minimizing the within-cluster sum of squared distances between observations and their corresponding centroids. Despite its popularity, the method requires validation procedures to ensure that the discovered clusters represent meaningful structures rather than artifacts of initialization or scaling choices.

II.3 CLUSTER VALIDATION AND AGREEMENT METRICS

Internal validation of clustering quality is commonly assessed using the Silhouette coefficient, which quantifies the relative separation between clusters by comparing intra-cluster cohesion and inter-cluster dissimilarity [9]. The coefficient ranges between -1 and 1 , with higher values indicating better-defined partitions. Beyond internal validation, external comparison between clustering solutions can be evaluated using the Adjusted Rand Index (ARI), which measures similarity between two partitions while correcting for chance agreement [10]. ARI values close to 1 indicate strong concordance between clustering methods, providing robust evidence for regime identification.

II.4 STATISTICAL VALIDATION OF REGIME SEPARABILITY

While clustering algorithms partition observations into groups, statistical inference is necessary to verify whether these groups differ significantly in terms of their defining characteristics. One-way Analysis of Variance (ANOVA) tests whether mean differences between groups are statistically significant [17]. However, statistical significance alone does not quantify the magnitude of group differences. Effect size metrics, such as eta squared (η^2), measure the proportion of total variance explained by the grouping factor [11]. Large η^2 values indicate substantial structural separation between regimes, thereby strengthening interpretability.

II.5 TEMPORAL STABILITY OF CLIMATE REGIMES

Climate classification over extended periods enables assessment of regime persistence and structural stability. Long-term hydro-climatic datasets make it possible to test whether identified regimes remain stable across historical sub-periods or undergo structural transitions [1]. Evaluating temporal stability enhances the scientific robustness of regionalization frameworks and provides insights into potential climate regime shifts under changing climatic conditions.

III. MATERIALS AND METHODS

III.1 DATASET DESCRIPTION

This study used monthly rainfall data from the Indian Meteorological Department (IMD), covering the period 1901–2015 across 36 meteorological subdivisions. The dataset provided long-term monthly precipitation totals (in millimeters), enabling the characterization of hydro-climatic behavior over 115 years. The temporal resolution was monthly, and the spatial aggregation level corresponded to climatologically defined subdivisions rather than administrative units. The extended historical coverage allowed assessment of long-term structural stability and variability patterns in rainfall regimes. The dataset was publicly available and was widely used in climate variability research [18]. However, the dataset consisted primarily of raw monthly rainfall observations, which did not directly encode higher-order structural characteristics of regional rainfall regimes. Consequently, additional climatic descriptors were constructed from the monthly time series in order to summarize annual magnitude, interannual variability, intra-annual seasonality, and rainfall concentration. These derived indicators provided a compact and physically interpretable representation of rainfall regimes and served as the basis for the clustering analysis described in the next section.

III.2 FEATURE ENGINEERING

To transform raw monthly rainfall observations into regime-oriented descriptors, a structured set of climatic indicators was derived from the rainfall time series. These descriptors were designed to summarize rainfall magnitude, variability, seasonal structure, and intra-annual concentration in a form suitable for unsupervised classification. Let $R_{i,t,m}$ denote the amount of rainfall (mm) observed in subdivision i , during year t , and month m , where:

- i indexed the meteorological subdivision,
- $t = 1, \dots, T$ indexed years,
- $m = 1, \dots, 12$ indexed months,
- $T = 115$ for the full study period (1901–2015).

III.2.1 Annual Magnitude Metrics

To characterize overall rainfall magnitude at the annual scale, monthly precipitation values were first aggregated into yearly totals:

$$A_{i,t} = \sum_{m=1}^{12} R_{i,t,m} \quad (1)$$

Where $A_{i,t}$ represented the annual rainfall (mm) for subdivision i in year t , obtained by summing the twelve monthly observations. This aggregation converted monthly measurements into a yearly climatic signal, enabling analysis of long-term interannual behavior. To quantify the long-term average rainfall magnitude of each subdivision, the mean annual rainfall was computed:

$$\mu_i = \frac{1}{T} \sum_{t=1}^T A_{i,t} \quad (2)$$

Where μ_i denoted the mean annual rainfall for subdivision i , averaged across the full study period. This metric captured the typical climatological rainfall level of each region. To measure interannual variability around the long-term mean, the standard deviation was calculated:

$$\sigma_i = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (A_{i,t} - \mu_i)^2} \quad (3)$$

Where σ_i quantified the dispersion of annual rainfall totals around μ_i . The denominator $T - 1$ ensured an unbiased estimate of variability. This indicator reflected the degree of climatic stability versus fluctuation across years. To normalize variability relative to rainfall magnitude and allow comparison between wet and dry regions, the coefficient of variation was computed:

$$CV_i = \frac{\sigma_i}{\mu_i} \quad (4)$$

provided that $\mu_i \neq 0$.

Here, CV_i was dimensionless and expressed relative variability independent of absolute rainfall magnitude. Together, these four metrics described the annual-scale structure of rainfall regimes in terms of magnitude and interannual variability [19].

III.2.2 Seasonal Structure and Intra-Annual Organization

While annual metrics captured long-term magnitude and variability, rainfall regimes are also defined by their intra-annual distribution. To characterize the seasonal cycle, monthly climatological means were computed:

$$\bar{R}_{i,m} = \frac{1}{T} \sum_{t=1}^T R_{i,t,m} \quad (5)$$

Where $\bar{R}_{i,m}$ represented the average rainfall in month m across all years for subdivision i . This calculation defined the typical seasonal rainfall profile of each subdivision. To identify the peak and trough of the seasonal cycle, the maximum and minimum climatological monthly means were determined:

$$R_i^{max} = \max_m(\bar{R}_{i,m}) \quad (6)$$

$$R_i^{min} = \min_m(\bar{R}_{i,m}) \quad (7)$$

Here, R_i^{max} corresponded to the wettest climatological month, while R_i^{min} represented the driest. These quantities identified the amplitude bounds of the annual rainfall cycle. To quantify the relative strength of seasonality, a normalized seasonal amplitude index was computed:

$$S_i = \frac{R_i^{max} - R_i^{min}}{\mu_i} \quad (8)$$

Where S_i measured the difference between the wettest and driest climatological months relative to the mean annual rainfall. Higher values indicated strongly seasonal regimes, whereas lower values suggested more evenly distributed rainfall. Finally, to evaluate rainfall concentration within a limited number of months, a concentration index was computed:

$$C_i = \frac{\sum_{k=1}^3 \bar{R}_{i,k}}{\sum_{m=1}^{12} \bar{R}_{i,m}} \quad (9)$$

Where the monthly climatological means were ranked in descending order and the three highest values were retained in the numerator. The denominator represented total climatological annual rainfall. This index quantified the proportion of rainfall concentrated in the three wettest months, thereby capturing the degree of intra-annual rainfall concentration. These descriptors capture intra-annual rainfall distribution patterns [20].

III.3 CLUSTERING PROCEDURE

To identify structurally distinct rainfall regimes, unsupervised clustering was applied to the derived feature matrix. Before clustering, all features were standardized using Z-score normalization to ensure comparability across variables with different units and scales:

$$Z_{i,j} = \frac{X_{i,j} - \bar{X}_j}{s_j} \quad (10)$$

Where:

- $X_{i,j}$ represented the original value of feature j for subdivision i ,
- $\bar{X}_j$ denoted the mean of feature j across all subdivisions,
- s_j was the standard deviation of feature j .

This normalization centered each feature at zero and scaled it to unit variance, preventing variables with larger numerical ranges from dominating the clustering process. K-Means clustering was applied to partition the standardized dataset homogeneous groups. A small p-value indicated that the observed difference in feature means between clusters were unlikely to have occurred by chance. The method aimed to minimize the within-cluster sum of squared Euclidean distances:

$$J = \sum_{k=1}^K \sum_{x \in C_k} |x - \mu_k|^2 \quad (11)$$

Where:

- C_k represented cluster k ,
- μ_k denoted the centroid of cluster k ,
- x corresponded to the feature vector of a subdivision.

This optimization assigned subdivisions to clusters such that within-cluster similarity was maximized and between-cluster separation was enhanced. The optimal number of clusters k was determined using the Silhouette coefficient [9], which evaluated cohesion and separation simultaneously. The number of clusters that maximized the average Silhouette score was selected. The Silhouette analysis indicated that $k = 2$ produced the highest average Silhouette coefficient among the tested values. This result suggested that a two-cluster partition maximized intra-cluster cohesion while maintaining clear inter-cluster separation. Increasing the number of clusters did not substantially improve the Silhouette score and led to smaller, less stable groups with reduced interpretability. Therefore, $k = 2$ was retained as the optimal solution, balancing statistical validity, structural robustness, and climatological interpretability. K-Means was implemented in Python using the scikit-learn library, with $n_{init} = 100$ and $random_{state} = 42$ to ensure solution stability and reproducibility.

III.4 ROBUSTNESS ASSESSMENT

To evaluate whether the identified rainfall regimes were algorithm-dependent, Ward hierarchical clustering was applied to the same standardized feature space used for K-Means clustering. Ward's method minimized total within-cluster variance at each agglomerative step, producing a hierarchical partition of subdivisions based on variance reduction criteria [21]. Agreement between K-Means and hierarchical clustering partitions was quantified using the Adjusted Rand Index (ARI) [10]:

$$ARI = \frac{Index - Expected}{Maximum - Expected} \quad (12)$$

Where:

- *Index* denoted the observed agreement between the two clustering solutions, measured in terms of pairwise assignments of subdivisions.
- *Expected* represented the level of agreement expected by chance under a random partitioning model with the same cluster sizes.
- *Maximum* corresponded to the maximum possible agreement between the two partitions.

The ARI therefore corrected the raw agreement measure for chance effects, ensuring that a value of 0 indicated agreement no better than random assignment, while a value of 1 indicated perfect concordance between clustering solutions. The number of clusters k was fixed to the optimal value identified in the full-period analysis to ensure methodological consistency and direct comparability between clustering approaches. A high ARI value supported the structural robustness of the identified rainfall regimes, indicating that the detected clustering structure was not an artifact of a specific algorithmic choice.

III.5 STATISTICAL VALIDATION

To statistically validate the structural separation between the identified rainfall regimes, one-way Analysis of Variance (ANOVA) was performed on each derived climatic descriptor. While clustering grouped subdivisions based on multivariate similarity, ANOVA was applied to determine whether the mean values of individual features differed significantly across clusters. This step ensured that the identified regimes were not only algorithmically distinct but also statistically separable.

For each feature, the ANOVA F-statistic was computed as:

$$F = \frac{MS_{between}}{MS_{within}} \quad (13)$$

Where:

- $MS_{between}$ represented the mean square between clusters, reflecting variability of feature means across regimes [17].
- MS_{within} denoted the mean square within clusters, reflecting variability among subdivisions belonging to the same regime [17].

The F-statistic therefore measured the ratio of inter-cluster variance to intra-cluster variance. Large F values indicated stronger separation between clusters relative to internal variability. Statistical significance was evaluated using the associated p-values. A small p-value indicated that the observed differences in feature means between clusters were unlikely to have occurred by chance. Because statistical significance alone may be influenced by sample size, effect size was additionally quantified using eta squared:

$$\eta^2 = \frac{SS_{between}}{SS_{total}} \quad (14)$$

where:

- $SS_{between}$ represented the between-cluster sum of squares.
- SS_{total} denoted the total sum of squares.

The coefficient η^2 measured the proportion of total variance in a feature that was explained by cluster membership. Higher η^2 values indicated stronger structural differentiation between rainfall regimes [11]. In contrast to p-values, which assessed statistical significance, eta squared quantified the magnitude of separation and provided an interpretable measure of practical relevance. Together, ANOVA and effect size analysis ensured that the identified rainfall regimes exhibited statistically significant and substantively meaningful differences across key climatic descriptors.

III.6 TEMPORAL STABILITY ANALYSIS

To evaluate the temporal persistence of the identified rainfall regimes, a stability analysis was conducted across distinct historical sub-periods. While the full-period clustering provided a global structural classification, it remained necessary to assess whether regime membership was temporally stable or sensitive to long-term climatic variability. The dataset was therefore divided into three historical sub-periods:

- 1901–1950
- 1951–1980
- 1981–2015

These intervals were selected to represent early, mid, and recent phases of the 20th and early 21st centuries, enabling assessment of long-term structural evolution. For each sub-period, the complete feature engineering procedure described in Section III.2 was repeated independently. Derived features were recalculated using only the data corresponding to the specific time window. Subsequently, feature standardization was performed separately within each sub-period to ensure internal comparability prior to clustering. K-Means clustering was then applied using the same number of clusters k identified in the full-period analysis. Fixing k ensured structural comparability across periods and prevented artificial differences arising from varying cluster numbers. A regime transition was defined as a change in cluster membership between consecutive sub-periods. Temporal stability was quantified as the proportion of subdivisions retaining identical cluster assignments across periods. Stability was evaluated for:

- consecutive transitions (1901–1950 vs 1951–1980; 1951–1980 vs 1981–2015),
- overall persistence across all three periods.

Subdivisions that remained in the same cluster across all periods were considered structurally stable, whereas subdivisions exhibiting at least one change were classified as transitional. This analysis allowed assessment of whether rainfall regimes represented persistent structural climatic patterns or whether they exhibited evidence of temporal reorganization.

IV. RESULTS AND DISCUSSIONS

This section presents the structural organization of Indian rainfall regimes derived from the clustering framework described in Section III. The analysis proceeds sequentially from the determination of the optimal number of clusters to their physical interpretation, methodological robustness, statistical separability, and temporal persistence. Together, these steps provide a comprehensive evaluation of the stability and validity of the identified rainfall regimes.

IV.1 OPTIMAL NUMBER OF RAINFALL REGIMES

The first step in the results analysis consisted of determining the optimal number of rainfall regimes represented in the standardized feature space. The clustering procedure conducted over the full historical period (1901–2015) indicated that $k = 2$ provided the most structurally coherent partition.

Among the tested configurations ($k = 2-6$), the two-cluster solution achieved the highest average Silhouette coefficient (0.428), reflecting the strongest balance between intra-cluster cohesion and inter-cluster separation (**Table 1**). Although the four-cluster configuration yielded a moderately close value (0.396), additional partitions primarily subdivided already coherent groups without generating physically distinct hydro-climatic structures. This fragmentation reduced interpretability without improving structural clarity. Consequently, the two-cluster configuration was retained as the most parsimonious and statistically justified solution.

Table 1: Silhouette scores for candidate numbers of clusters ($k = 2 \dots 6$) over 1901–2015.

Number of clusters (K)	Silhouette score
2	0.428
3	0.384
4	0.396
5	0.308
6	0.335

Source: Authors (2026).

Geometric confirmation of this partition was provided by the PCA projection (Figure 1). The two clusters were predominantly separated along the first principal component, which captured the dominant variance structure of the engineered climatic descriptors. Limited overlap was observed for a small subset of subdivisions, consistent with the moderate Silhouette value. This pattern suggests that rainfall regime differentiation follows gradual hydro-climatic gradients rather than abrupt discontinuities.

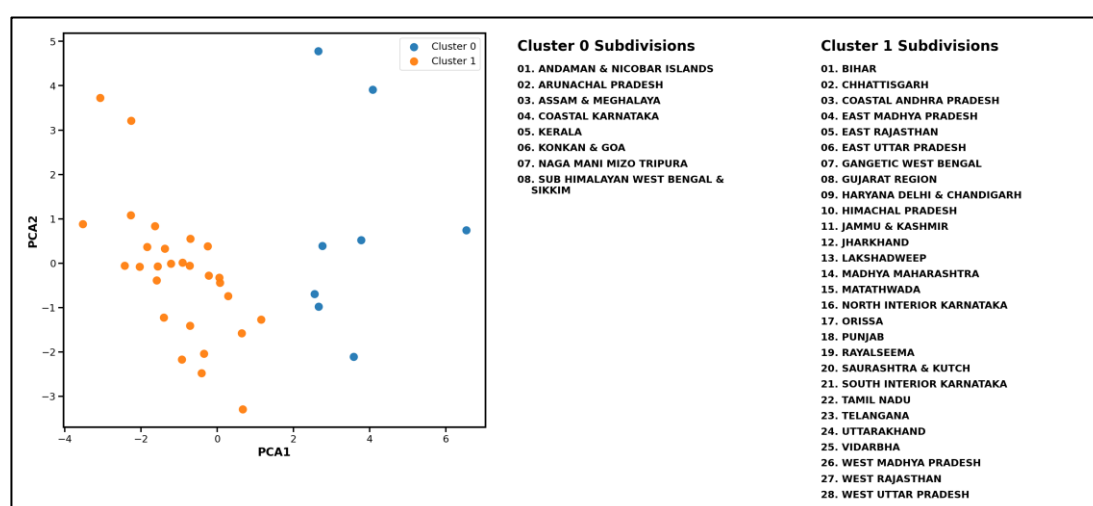


Figure 1: PCA projection of subdivisions using K-Means clustering ($k = 2$).

Source: Authors (2026).

Having established the two-cluster solution as statistically optimal and geometrically coherent, the next step involved interpreting the physical and climatic characteristics that distinguish these rainfall regimes. This interpretation is essential to determine whether the statistical partition corresponds to meaningful hydro-climatic structures rather than purely mathematical groupings.

IV.2 INTERPRETATION OF THE TWO RAINFALL REGIMES

The two identified clusters corresponded to statistically distinguishable and physically interpretable rainfall regimes primarily differentiated by rainfall magnitude and seasonal peak intensity.

• Regime A (Cluster 0: High-Rainfall / High-Peak Regime)

- Mean annual rainfall ≈ 2920.81 mm
- Maximum annual rainfall ≈ 4432.06 mm
- Maximum monthly climatological mean ≈ 708.67 mm

These values indicate persistently high annual precipitation totals combined with pronounced seasonal concentration. The large maximum monthly mean suggests strong monsoonal dominance, where a substantial proportion of annual rainfall is delivered within a limited number of peak months. Spatial inspection revealed that these subdivisions were predominantly located in coastal and northeastern regions, areas historically influenced by intense monsoon circulation and orographic enhancement. The coherence of magnitude-related variables across this cluster indicates a structurally wet hydro-climatic regime characterized by high water availability and strong seasonal peaks.

• Regime B (Cluster 1: Lower-Rainfall / Lower-Peak Regime)

- Mean annual rainfall ≈ 997.47 mm
- Maximum annual rainfall ≈ 1547.81 mm
- Maximum monthly mean ≈ 257.31 mm

Compared to Regime A, these values reflect substantially lower annual totals and weaker seasonal peaks. Rainfall in this regime was more moderate in magnitude and generally less concentrated in extreme peak months. These subdivisions corresponded largely to interior and semi-arid regions, where continental effects and reduced monsoonal penetration limit total precipitation. The multivariate shift observed in the standardized feature heatmap (Figure 2) confirms that regime differentiation was not driven by a single variable but by coherent changes across multiple magnitude-related descriptors.

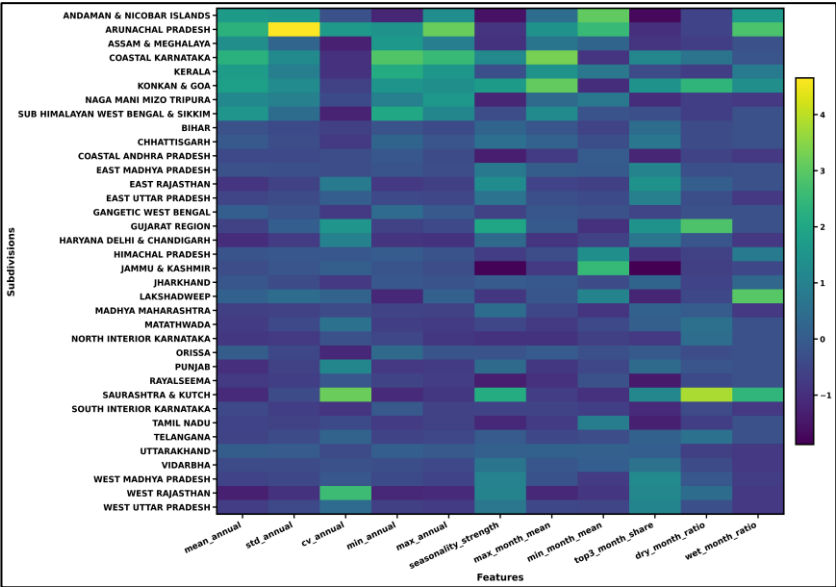


Figure 2: Standardized feature heatmap (subdivisions × features) sorted by K-Means cluster.
Source: Authors (2026).

While the two-regime structure exhibited clear physical interpretability and coherent multivariate differentiation, it remained necessary to verify that this partition was not dependent on the specific clustering algorithm employed. Establishing methodological robustness is essential to ensure that the identified rainfall regimes reflect an intrinsic organization of the feature space rather than an artifact of centroid initialization or optimization strategy.

IV.3 ROBUSTNESS OF THE CLUSTERING SOLUTION

To assess algorithmic robustness, Ward hierarchical clustering was applied to the same standardized feature matrix with the number of clusters fixed at $k = 2$. Ward’s method minimizes total within-cluster variance at each agglomerative step, providing an alternative variance-based partitioning framework. The hierarchical solution produced a comparable Silhouette coefficient (0.413) and demonstrated strong agreement with the K-Means partition, as quantified by an Adjusted Rand Index (ARI) of 0.881. An ARI value close to 1 indicates near-equivalent partitions after correcting for chance agreement. The obtained value suggests that the detected rainfall regimes were not artifacts of centroid-based optimization, but rather reflected a stable structural separation embedded in the multivariate rainfall feature space. The dendrogram representation (Figure 3) further illustrates that the dataset naturally separated into two dominant branches consistent with the K-Means solution.

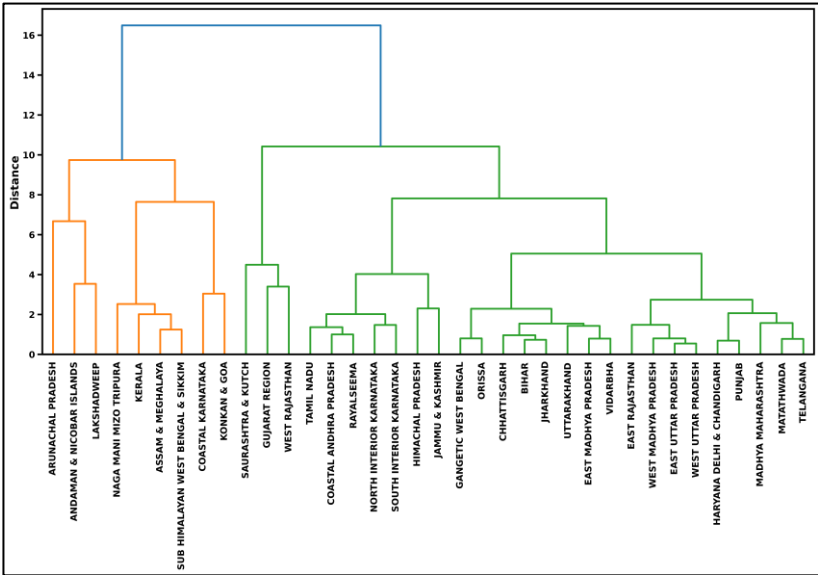


Figure 3: Ward hierarchical dendrogram ($k = 2$, consistent with K-Means).
Source: Authors (2026).

Although the robustness analysis confirmed that the two-regime structure was not algorithm-dependent, it remained necessary to determine whether the identified clusters were statistically distinguishable in terms of their underlying climatic descriptors. Structural robustness alone does not guarantee that regime separation reflects meaningful differences in feature distributions. Therefore, inferential statistical testing was performed to formally evaluate regime separability.

IV.4 STATISTICAL VALIDATION OF REGIME SEPARABILITY

To formally assess whether the two rainfall regimes differed significantly across the engineered climatic descriptors, one-way Analysis of Variance (ANOVA) was applied to each feature. While clustering identified multivariate groupings based on similarity, ANOVA evaluated whether the mean values of individual descriptors differed significantly between regimes. This step ensured that the separation observed in the clustering space corresponded to statistically measurable differences in rainfall characteristics rather than arbitrary partitioning. Several rainfall descriptors exhibited highly significant differences ($p < 0.001$) accompanied by large effect sizes (η^2), indicating substantial structural differentiation between regimes. The strongest discriminative variables were:

- **mean_annual:** $F = 208.10$, $p < 0.001$, $\eta^2 = 0.860$
- **max_annual:** $F = 162.53$, $p < 0.001$, $\eta^2 = 0.827$
- **max_month_mean:** $F = 63.70$, $p < 0.001$, $\eta^2 = 0.652$
- **min_annual:** $F = 47.81$, $p < 0.001$, $\eta^2 = 0.584$
- **std_annual:** $F = 45.30$, $p < 0.001$, $\eta^2 = 0.571$

Effect sizes exceeding 0.5 indicate that more than half of the variance in the respective descriptor was explained by regime membership. In particular, mean_annual and max_annual exhibited η^2 values above 0.80, demonstrating that cluster membership accounted for the majority of variability in annual rainfall magnitude. In contrast, relative variability indicators such as the coefficient of variation exhibited weaker statistical separation ($p = 0.1138$), suggesting that absolute rainfall magnitude played a more central role in defining hydro-climatic regimes at the subdivision scale than relative dispersion metrics. Table 2 summarizes the ANOVA statistics and corresponding effect sizes, ranked by η^2 . The dominance of magnitude-related features confirms that regime differentiation is structurally anchored in total precipitation availability and seasonal peak intensity rather than minor fluctuations.

Table 2: ANOVA results and effect sizes (η^2) for engineered rainfall descriptors (sorted by η^2).

Feature	F-statistic	p-value	η^2
mean_annual	208.10	< 0.001	0.860
max_annual	162.53	< 0.001	0.827
max_month_mean	63.70	< 0.001	0.652
min_annual	47.81	< 0.001	0.584
std_annual	45.30	< 0.001	0.571

Source: Authors (2026).

Figure 4 provides a complementary distributional perspective of the strongest discriminative features. The boxplots show clear separation of median values between regimes, with limited interquartile overlap for magnitude-related descriptors. The persistence of separation across the full range of observed values indicates that regime differentiation is structural rather than driven by isolated extreme observations. This visual confirmation reinforces the inferential results and supports the robustness of the identified rainfall regimes.

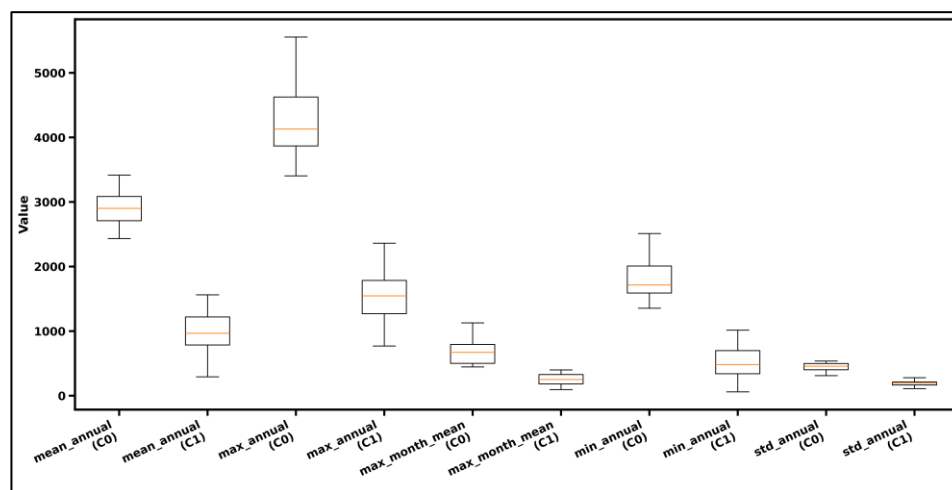


Figure 4: Boxplots of the strongest discriminative features between clusters (Top η^2).

Source: Authors (2026).

Although the inferential analysis confirmed that the two regimes were statistically separable across key climatic descriptors, it remained necessary to examine whether this regime structure was temporally persistent. Statistical separability over the full period does not, by itself, guarantee that regime membership was stable under long-term climatic variability. Therefore, a temporal stability analysis was conducted by repeating the full feature engineering and clustering procedure over distinct historical sub-periods.

IV.5 TEMPORAL STABILITY ANALYSIS

To evaluate whether the identified rainfall regimes represented persistent hydro-climatic structures rather than period-specific artifacts, a temporal stability analysis was conducted across three historical sub-periods: 1901–1950, 1951–1980, and 1981–2015. For each sub-period, the complete feature engineering pipeline described in Section III.2 was recomputed using only data from the corresponding time window. Features were standardized independently within each period to preserve internal comparability. K-Means clustering was then applied with the number of clusters fixed at $k = 2$, as determined in the full-period analysis, ensuring structural consistency across temporal partitions. The Silhouette coefficients obtained for the three sub-periods were:

- 1901–1950: Silhouette = 0.423
- 1951–1980: Silhouette = 0.456
- 1981–2015: Silhouette = 0.463

The stability of Silhouette values across periods indicates that regime separability remained structurally coherent over time. Notably, cluster separation slightly increased in the more recent periods, suggesting that hydro-climatic differentiation did not weaken under long-term variability. To quantify regime persistence at the subdivision level, cluster membership was tracked across consecutive periods. A regime transition was defined as a change in cluster assignment between adjacent sub-periods. Out of the 36 meteorological subdivisions, 35 retained identical cluster membership across all three periods. Only one subdivision (Lakshadweep) exhibited a transition between the first and second sub-periods, after which its cluster membership remained stable. Table 3 reports regime membership by sub-period and identifies detected transitions.

Table 3: Regime membership by period and detected transitions (1901–1950, 1951–1980, 1981–2015).

Subdivision	P1	P2	P3	P1 → P2	P2 → P3	Any transition
Lakshadweep	0	1	1	1	0	1

Source: Authors (2026).

The near-complete persistence of regime assignments indicates that the two-cluster structure reflected a stable large-scale hydro-climatic organization rather than transient fluctuations. The limited occurrence of transitions further suggests that regime boundaries were robust to multi-decadal climatic variability. Figure 5 summarizes this stability pattern by contrasting unchanged and changed subdivisions.

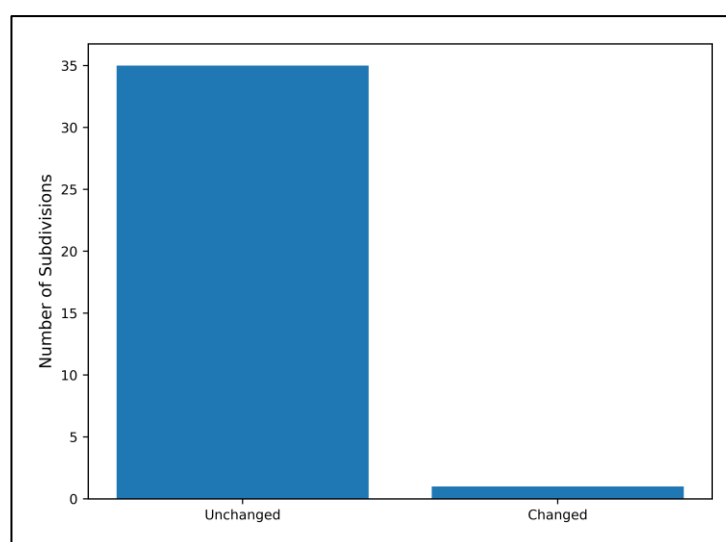


Figure 5: Regime stability across historical periods (unchanged vs. changed subdivisions).

Source: Authors (2026).

The dominance of unchanged assignments across more than a century reinforces the interpretation that Indian rainfall organization can be effectively described by two temporally persistent regimes. While interannual variability is well documented in monsoon dynamics, the underlying structural partition of subdivisions remained structurally stable throughout the study period. Having demonstrated statistical separability, algorithmic robustness, and temporal persistence, the identified rainfall regimes can now be interpreted in a broader scientific context. The implications of this structural organization are discussed below, followed by a clarification of the analytical scope of the present study.

IV.6 SCIENTIFIC IMPLICATIONS AND ANALYTICAL SCOPE

The results indicate that long-term Indian rainfall organization at the subdivision scale can be consistently represented by two statistically coherent and temporally stable hydro-climatic regimes. The convergence of internal validation metrics, cross-method agreement, and inferential testing supports the structural validity of this partition. From a scientific standpoint, the identification of two persistent regimes provides a parsimonious regionalization framework for large-scale hydro-climatic assessment. By reducing a high-dimensional rainfall dataset into interpretable regime categories, the analysis facilitates comparative evaluation across subdivisions and offers a structured basis for examining spatial patterns of water availability and seasonal intensity.

It is important to clarify that the present study focused specifically on long-term rainfall magnitude and seasonal organization derived from monthly totals. The objective was not to construct a comprehensive climatic zonation, but rather to evaluate whether stable structural rainfall regimes could be identified using interpretable precipitation descriptors. Accordingly, the regime classification should be understood as a rainfall-centered structural framework at the meteorological subdivision scale. Future extensions may incorporate additional climatic variables or alternative temporal resolutions to assess how the regime structure evolves under expanded feature representations.

V. CONCLUSIONS

This study investigated the large-scale structural organization of Indian rainfall using a long-term monthly dataset spanning 1901–2015 and covering 36 meteorological subdivisions. By transforming raw precipitation observations into interpretable climatic descriptors and applying unsupervised clustering techniques, the analysis identified two coherent rainfall regimes that consistently characterized subdivision-level hydro-climatic behavior. The two-regime structure was supported by multiple layers of validation. Internal cohesion and separation metrics indicated an optimal partition at $k = 2$, geometric projection confirmed clear structural differentiation, cross-method comparison demonstrated algorithmic robustness, inferential testing established statistical separability, and temporal replication showed persistent regime membership across historical sub-periods. Taken together, these complementary evaluations provide convergent evidence that the identified regimes reflect a stable organization embedded in the multivariate rainfall feature space. The results suggest that, at the subdivision scale, long-term Indian rainfall variability can be meaningfully summarized through a parsimonious two-regime framework primarily distinguished by differences in rainfall magnitude and seasonal intensity. This structural simplification does not aim to replace detailed regional climatology but offers an interpretable representation suitable for comparative assessment and large-scale hydro-climatic characterization. It is important to emphasize that the present analysis focused on rainfall-derived descriptors at monthly resolution and subdivision scale. Accordingly, the findings should be interpreted within this analytical scope. Extensions incorporating additional climatic variables, higher spatial resolution data, or event-based indices may further refine the understanding of rainfall regime dynamics. Overall, the study demonstrates that a statistically grounded and temporally persistent regime structure underlies Indian rainfall organization over more than a century, providing a robust foundation for future investigations of hydro-climatic variability and regional water-resource assessment.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Imane Meguenni and Bendaoud Mebarek.

Methodology: Imane Meguenni and Bendaoud Mebarek.

Investigation: Imane Meguenni and Bendaoud Mebarek.

Discussion of results: Imane Meguenni and Bendaoud Mebarek.

Writing – Original Draft: Imane Meguenni and Bendaoud Mebarek.

Writing – Review and Editing: Imane Meguenni and Bendaoud Mebarek.

Resources: Imane Meguenni and Bendaoud Mebarek.

Supervision: Bendaoud Mebarek.

Approval of the final text: Bendaoud Mebarek.

VII. REFERENCES

- [1] V. Masson-Delmotte et al., Eds., *Climate Change 2021: The Physical Science Basis: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, U.K., and New York, NY, USA: Cambridge University Press, 2021. doi: 10.1017/9781009157896.
- [2] Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2022: Impacts, Adaptation and Vulnerability: Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, U.K.: Cambridge University Press, 2023. doi: 10.1017/9781009325844.
- [3] R. Krishnan, J. Sanjay, C. Gnanaseelan, M. Mujumdar, A. Kulkarni, and S. Chakraborty, Eds., *Assessment of Climate Change over the Indian Region: A Report of the Ministry of Earth Sciences (MoES)*, Government of India, 1st ed. Singapore: Springer, 2020. doi: 10.1007/978-981-15-4327-2.
- [4] R. Huth, C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynová, J. Kyselý, and O. E. Tveito, "Classifications of atmospheric circulation patterns: Recent advances and applications," *Annals of the New York Academy of Sciences*, vol. 1146, no. 1, pp. 105–152, 2008. doi: 10.1196/annals.1446.019.
- [5] D. Radinović and M. Ćurić, "Measuring scales for daily temperature extremes, precipitation and wind velocity," *Meteorological Applications*, vol. 21, no. 3, pp. 461–465, 2014. doi: 10.1002/met.1356.
- [6] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al., "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. doi: 10.1002/qj.3803.
- [7] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. doi: 10.1016/j.patrec.2009.09.011.
- [8] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. doi: 10.1109/TIT.1982.1056489.
- [9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.
- [10] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985. doi: 10.1007/BF01908075.
- [11] S. Olejnik and J. Algina, "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," *Psychological Methods*, vol. 8, no. 4, pp. 434–447, 2003. doi: 10.1037/1082-989X.8.4.434.
- [12] E. Xoplaki, J. F. González-Rouco, J. Luterbacher, et al., "Wet season Mediterranean precipitation variability: Influence of large-scale dynamics and trends," *Climate Dynamics*, vol. 23, pp. 63–78, 2004. doi: 10.1007/s00382-004-0422-0.

- [13] FAO, The State of Food and Agriculture 2022: Leveraging Automation in Agriculture for Transforming Agrifood Systems. Rome, Italy: Food and Agriculture Organization of the United Nations, 2022. doi: 10.4060/cb9479en.
- [14] B. Yarnal, "Synoptic Climatology in Environmental Analysis: A Primer," Environmental Impact Assessment Review, vol. 14, no. 1, pp. 83–84, 1994. doi: 10.1016/0160-4120(93)90293-Q.
- [15] V. Eyring, W. D. Collins, P. Gentine, E. A. Barnes, M. Barreiro, T. Beucler, M. Bocquet, C. S. Bretherton, H. M. Christensen, K. Dagon, et al., "Pushing the frontiers in climate modelling and analysis with machine learning," Nature Climate Change, vol. 14, no. 9, pp. 916–928, 2024. doi: 10.1038/s41558-024-02095-y.
- [16] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979. doi: 10.2307/2346830.
- [17] A. Field, Discovering Statistics Using IBM SPSS Statistics. London, U.K.: Sage Publications Ltd., 2024.
- [18] R. Ilangoan, "Rainfall in India: Sub-division wise monthly data for 115 years from 1901–2015," Kaggle Dataset, 2017. [Online]. Available: <https://www.kaggle.com/datasets/rajanand/rainfall-in-india>. Accessed: 01-Feb-2026.
- [19] D. S. Wilks, Statistical Methods in the Atmospheric Sciences, 4th ed. Amsterdam, The Netherlands: Elsevier, 2020. doi: 10.1016/C2017-0-03921-6.
- [20] R. P. D. Walsh and D. M. Lawler, "Rainfall seasonality: Description, spatial patterns and change through time," Weather, vol. 36, no. 7, pp. 201–208, 1981. doi: 10.1002/j.1477-8696.1981.tb05400.x.
- [21] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," Journal of the American Statistical Association, vol. 58, no. 301, pp. 236–244, 1963. doi: 10.1080/01621459.1963.10500845.