Report for end-semester evaluation of CE 498 course

Clustering Methods for the Analysis of Spatio-Temporal Distribution of Precipitation and Temperature Data in Northeast India

Submitted

By

Rohan Kumar Ishwar

Under the supervision of

Dr. Sreeja Pekkat

Department of Civil Engineering

Indian Institute of Technology Guwahati

November 2024

CERTIFICATE

It is certified that the work contained in the project report entitled " **Clustering Methods for the Analysis of Spatio-Temporal Distribution of Precipitation and Temperature Data in Northeast India** ", by **Rohan Kumar Ishwar** (210104089) has been carried out under my/our supervision and that this work has not been submitted elsewhere for the award of a degree or diploma.

Date: 27-11-2024

Signature
Dr. Sreeja Pekkat
Department of Civil Engineering
Indian Institute of Technology Guwahati

ABSTRACT

This study explores the application of advanced clustering techniques—K-Means, DBSCAN, and Hierarchical Clustering—to analyze the spatio-temporal distribution of precipitation and temperature in Northeast India, with the objective of uncovering regional climate patterns and trends. Each method is evaluated for its efficacy in managing spatial heterogeneity, temporal variability, and noise. Comparative analysis highlights DBSCAN's strength in detecting irregular and non-linear patterns, K-Means' efficiency in identifying structured clusters, and Hierarchical Clustering's capacity to reveal nested relationships. The findings underscore the potential of clustering to identify precipitation hotspots and temperature gradients, providing actionable insights for climate adaptation strategies, resource planning, and disaster mitigation. The report concludes by emphasizing the practical applicability of clustering techniques for analyzing complex spatio-temporal climatic data and outlines plans for further analysis in the next phase.

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to my supervisor, Dr. Sreeja Pekkat, Associate Professor, Department of Civil Engineering, Indian Institute of Technology Guwahati, for her invaluable guidance, insightful suggestions, and unwavering encouragement throughout the progression of this study and the preparation of this thesis.

My sincere thanks also go to Abdul Rahman (PhD Scholar) for his support, positive outlook on this work, and invaluable feedback throughout the course of this project.

I am deeply appreciative of their contributions and thank them wholeheartedly.

Date: 27-11-2024 **Rohan Kumar Ishwar**

TABLE OF CONTENTS

List of Figures

Chapter 1

INTRODUCTION

1.1 General

Climate variability and change have become critical global concerns due to their far-reaching impacts on ecosystems, economies, and human livelihoods. The Northeast region of India, renowned for its complex topography and diverse climatic conditions, is particularly vulnerable to climate-induced challenges. This region has witnessed a significant rise in extreme weather events, including heavy precipitation, floods, and temperature fluctuations, which have resulted in considerable socio-economic disruptions. According to recent reports, India experienced a 26% increase in extreme rainfall events between 2001 and 2020, with Northeast India contributing a substantial share. Such events exacerbate issues related to agriculture, water resource management, and disaster preparedness, underscoring the urgency of understanding regional climatic patterns.

1.2 Study area

Northeast region of India(as shown in fig.1) which consists of the eight states Assam, Meghalaya, Arunachal Pradesh, Tripura, Nagaland, Mizoram, Manipur and Sikkim is selected for the study. This region extends from latitude 22.4°N to 28.7°N and longitude 88.2°E to 96.5°E. Elevation varies from 28 m above mean sea level to 7000 m above mean sea level. Region is very diverse in nature, and plains mainly comprise Brahmaputra and Barak valleys. Onset of monsoon occurs from middle of May and continues till October. On average, the NE region receives about 2450 mm of rainfall. The Cherrapunjee–Mawsynram range receives rainfall as high as 11,500 mm, about 60% area under forest with Arunachal Pradesh hav ing about 80% of its area under different kinds of forest, while Assam has the minimum percentage of forest area (30%). Large altitude differences and varying physical features are the main reason of diverse climate from near tropical to temperate. The annual rainfall in the region is received mainly from the southwest monsoon, annually (Das et al. 2009; Dash et al. 2012; Goyal 2014). Northeast India is one of the most vulnerable regions toward climate change, and the largest threat to its biodiversity and hydrological system is from changing climate (Ravindranath et al. 2011). For studying the extreme precipitation indices and its trend, 27 stations from 0.5° × 0.5° gridded precipitation and temperature data from India Meteorological Department (IMD) were selected for the period of 1965–2014.

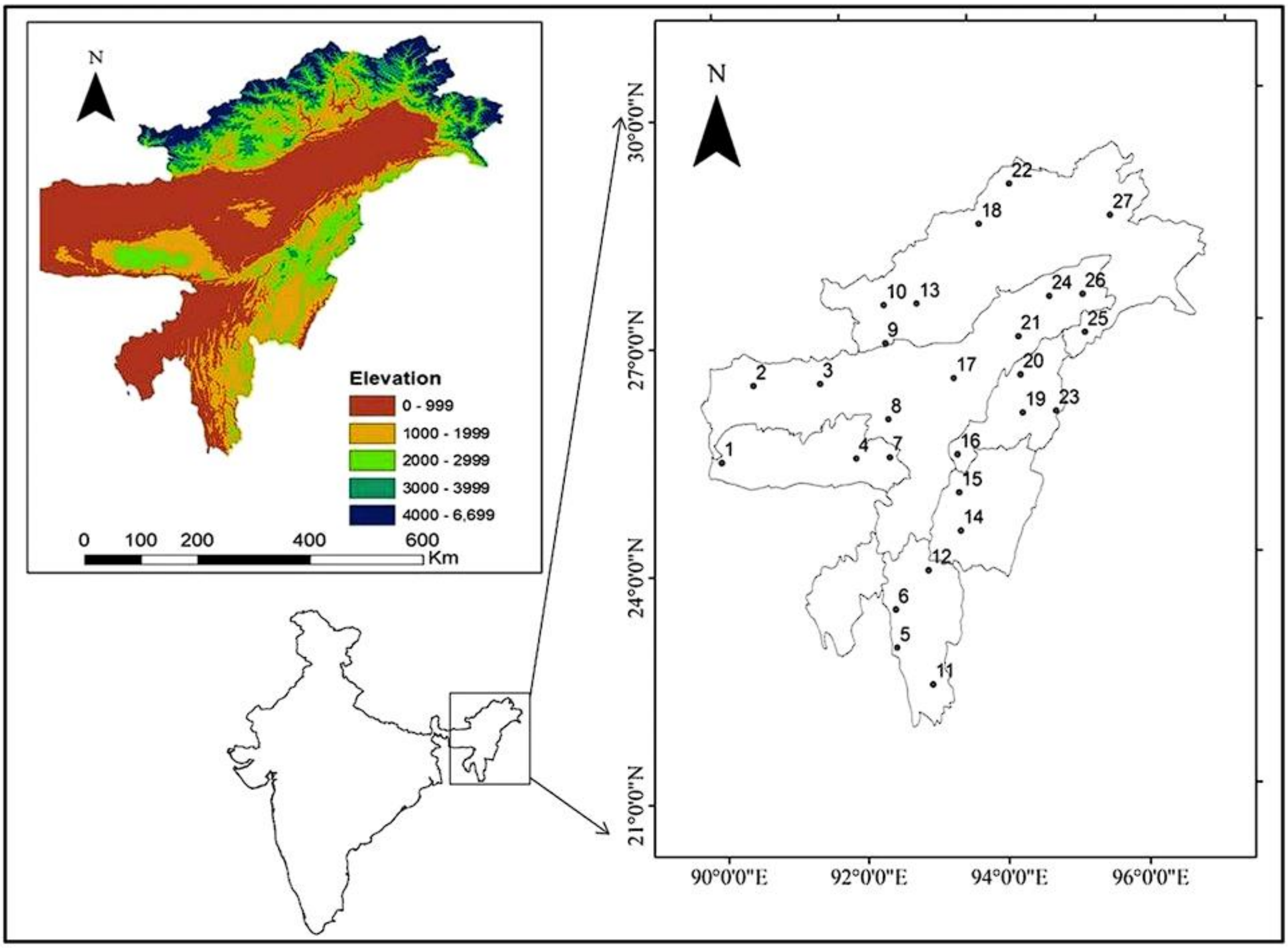


Fig. 1 Study area and location of selected station used to make gridded dataset

1.3 Climate Data

The increasing availability of high-resolution climatic datasets, driven by advancements in satellite observations and ground-based measurements, has revolutionized the field of climate science. These datasets offer unprecedented opportunities to analyse complex climate dynamics at finer spatial and temporal scales. However, the inherent complexity and vast volume of such datasets present significant analytical challenges. Extracting actionable insights requires methodologies capable of capturing subtle spatial variability and dynamic temporal interactions, which are often overlooked by traditional statistical approaches. The gridded data helped us analyse the trends using Mann–Kendall nonparametric test (Mann 1945; Kendall 1948) was used for calculating the trends of precipitation indices.

1.4 Clustering Analysis

Clustering analysis is an unsupervised learning technology, which can discover the structure and other information contained in the data itself without knowing the correct results in advance (Amin et al., 2020; Kar et al., 2015; Zhou et al., 2020). There are three common clustering methods: K-means clustering, hierarchical clustering and Density-based Spatial Clustering of Application with Noise (DBSACN) (Long et al., 2020; XIONG et al., 2020; Zheng et al., 2020). This approach facilitates the segmentation of datasets into meaningful clusters, which can reveal hidden relationships and regional characteristics that are critical for understanding climate variability and change.

By leveraging clustering methods such as K-Means, DBSCAN, and Hierarchical Clustering, it is possible to delineate climate zones, identify hotspots of extreme weather, and track seasonal transitions with greater precision. These insights are invaluable for addressing pressing challenges in climate-sensitive regions, such as optimizing agricultural practices, managing water resources, and improving disaster preparedness. The integration of clustering with high-resolution data analytics represents a significant step toward informed decision-making, enabling policymakers and researchers to better understand and mitigate the impacts of climate variability on ecosystems and human systems.

1.5 Motivation for work

The region's intricate topography and diverse climatic systems necessitate advanced analytical approaches to decipher spatio-temporal climate dynamics. With the advent of high-resolution datasets from satellite observations and ground-based systems, traditional statistical techniques fail to fully exploit the multidimensional complexity of the data. By deploying state-of-the-art clustering algorithms, this study seeks to uncover latent patterns, delineate precipitation anomalies, and identify temperature gradients, thereby enabling precise climate modeling and enhanced regional adaptive strategies for disaster risk reduction and sustainable resource management.

Chapter 2

LITERATURE REVIEW

2.1 Introduction to Climate Data and Spatio-Temporal Analysis

The study of precipitation and temperature distributions requires high-resolution spatio-temporal datasets to capture complex climatic variations. Advances in remote sensing technologies and ground-based observation networks have significantly improved the granularity of climate datasets, providing robust inputs for analytical frameworks. However, deriving actionable insights from these datasets poses challenges due to their sheer volume and dimensionality. This necessitates the use of machine learning approaches like clustering to extract patterns and trends from the data.

2.2 Datasets for Climate Analysis

2.2.1 Sources and Characteristics

Climate studies often rely on datasets from sources such as:

- **Satellite Observations**: Datasets like TRMM (Tropical Rainfall Measuring Mission) and GPM (Global Precipitation Measurement) offer high temporal and spatial resolution for precipitation data.

- **Ground-Based Stations**: Meteorological station networks provide localized and accurate measurements of precipitation and temperature.

- **Reanalysis Data**: Products like ERA5 and MERRA2 integrate observational data with model outputs for comprehensive climatic datasets.

These datasets are characterized by high dimensionality, temporal continuity, and spatial variability, requiring advanced computational techniques for analysis. The complexity of these characteristics—interdependencies across time, space, and variables—necessitates advanced computational techniques capable of handling the intricacies of spatial-temporal interactions.

2.2.2 Challenges in Dataset Utilization

- **Heterogeneity**: Integration of data from multiple sources introduces inconsistencies.
- **Missing Values**: Incomplete observations often hinder robust analysis.
- **Scale Variability**: Differences in spatial and temporal resolutions affect data harmonization.

2.2.3 Data Preprocessing

Outline pre-processing steps, such as handling missing values, standardizing data, or aggregating temporal resolutions (e.g., daily to monthly). By systematically addressing the preprocessing steps, the dataset becomes well-prepared for clustering, ensuring robust and interpretable results for spatio-temporal climate analysis.

2.3 Clustering Techniques

Traditional statistical approaches often fall short in addressing such multidimensional interdependencies, necessitating the use of machine learning methods like clustering. Clustering techniques are pivotal in identifying climate zones, precipitation hotspots, and temperature gradients. The following subsections review the clustering methods applied in this study:

2.3.1 K-Means Clustering

- **Overview**: K-Means is a partition-based clustering method that minimizes intra-cluster variance. It is widely used for structured datasets due to its computational efficiency.
- **Applications in Climate Studies**: Studies have demonstrated the utility of K-Means in classifying precipitation zones and temperature regimes. For example, K-Means has been employed to cluster rainfall patterns across monsoonal regions.
- **Limitations**: The method assumes spherical cluster shapes and requires the pre-specification of the number of clusters (k), making it less effective for irregularly shaped climate zones.

2.3.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Overview**: DBSCAN identifies clusters based on density, making it effective for datasets with noise and irregular cluster shapes.

- **Applications in Climate Studies**: DBSCAN excels in detecting precipitation anomalies and extreme events. Its ability to classify noise points is particularly valuable for identifying outliers in spatio-temporal climate data.
- **Limitations**: DBSCAN's performance is sensitive to parameters like epsilon (eps) and minimum samples, which require careful tuning based on data characteristics.

2.3.3 Hierarchical Clustering

- **Overview**: Hierarchical clustering builds a dendrogram to represent nested clusters, offering a flexible approach to cluster analysis without predefined cluster numbers.
- **Applications in Climate Studies**: Hierarchical clustering has been used to analyse nested climate patterns, such as sub-regional rainfall behaviours and temperature gradients.
- **Limitations:** The method is computationally intensive and sensitive to distance metrics and linkage methods, which can affect the resulting cluster structure**.**

2.3.4 Comparative Insights from Literature

The literature underscores the strengths and applications of various clustering methods in spatio-temporal climate analysis:

- **K-Means:** Effective for structured, low-noise datasets with well-defined clusters.
- **DBSCAN:** Robust for identifying irregular clusters and handling noise, particularly useful in detecting extreme climatic events.
- **Hierarchical Clustering:** Suitable for capturing nested relationships and multi-scale patterns, often applied in exploratory climate research.

The selection of a clustering method is influenced by dataset characteristics, such as noise levels, spatial variability, and temporal dynamics, as well as the specific analytical goals. A combined approach leveraging these methods provides a holistic understanding of spatio-temporal climatic patterns, enhancing the accuracy and depth of insights.

Chapter 3

Methodology

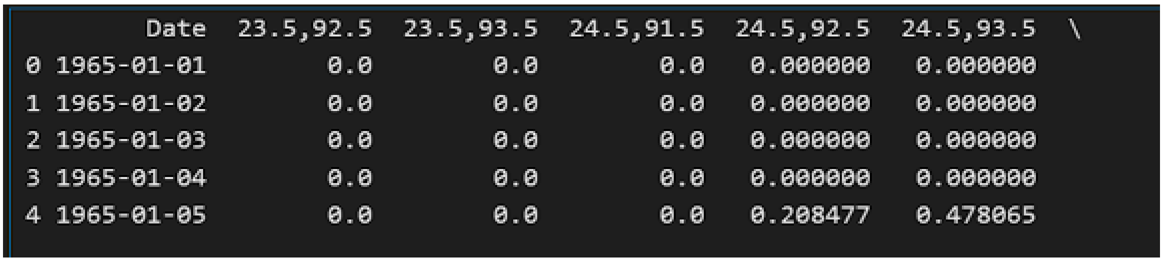
3.1. Data Preprocessing

3.1.1 Handling Missing Values

The preprocessing of climatic datasets (as shown in fig.2) involved addressing data(daily mean precipitation in mm) gaps caused by sensor malfunctions, poor connectivity, or data loss during transmission. Missing data were treated using:

- **Temporal Interpolation:** Suitable for filling gaps in continuous time-series data to preserve temporal trends.
- **Grid-wise Mean Imputation:** Addressed spatial gaps by averaging precipitation or temperature values across nearby grid points, minimizing spatial distortion.

The combination of these methods maintained the dataset's integrity while ensuring reliable input for downstream clustering analysis.



```
         Date  23.5,92.5  23.5,93.5  24.5,91.5  24.5,92.5  24.5,93.5  \
0 1965-01-01        0.0        0.0        0.0   0.000000   0.000000
1 1965-01-02        0.0        0.0        0.0   0.000000   0.000000
2 1965-01-03        0.0        0.0        0.0   0.000000   0.000000
3 1965-01-04        0.0        0.0        0.0   0.000000   0.000000
4 1965-01-05        0.0        0.0        0.0   0.208477   0.478065
```

Fig. 2 Sample Dataset Representation: Extracted Precipitation Metrics

3.1.2 Temporal Aggregation

To simplify the analysis of climatic patterns, daily data were aggregated into:

- **Monthly Means:** Highlighted seasonal variability, particularly useful for detecting monsoonal behaviour in the dataset.

- **Seasonal Aggregation:** Divided data into three climatic seasons (winter, monsoon, and summer), allowing a high-level analysis of intra-annual variability. For easy handling of data we divided each season having 4 months as shown in fig. 3.
    - Winter- November to February
    - Summer- March to June
    - Monsoon- July to October

3.1.3 Standardization

Climatic variables displayed large variability due to differences in precipitation intensity and temperature ranges across regions. Z-score normalization was applied to:

- Reduce biases caused by extreme values.
- Improve clustering performance by ensuring equal weighting of variables across all grid points.

This preprocessing laid the foundation for accurate clustering and reliable trend analysis. Finally, we had daily, monthly and seasonal data with corresponding latitude and longitude after this step.

3.2. Trend Analysis

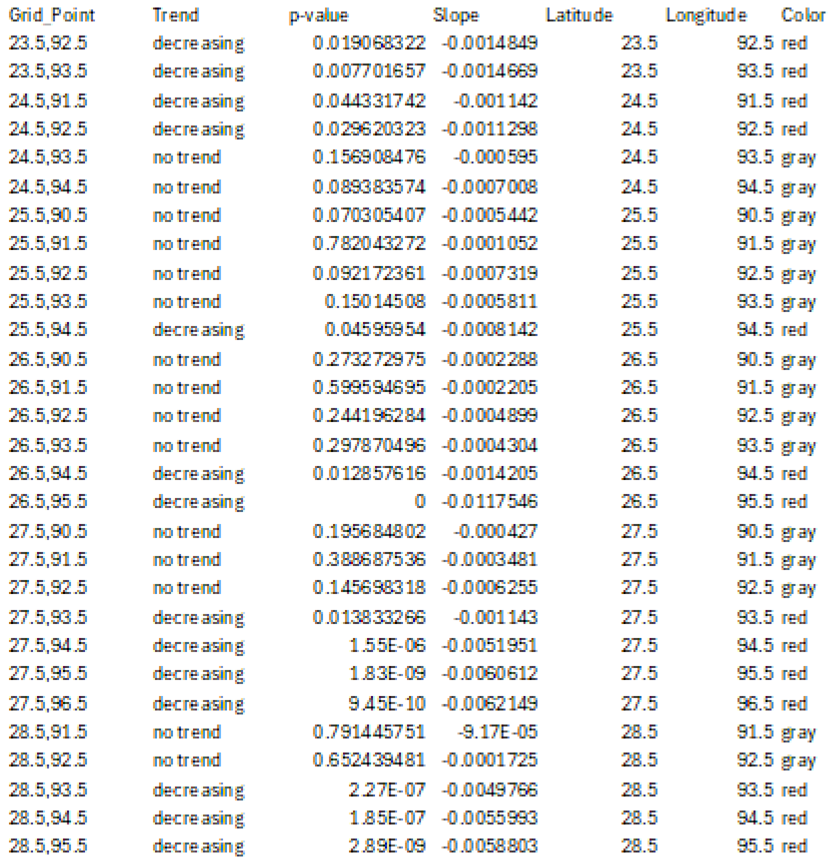
3.2.1 Mann-Kendall Test and Sen's Slope Estimator

Significant Trends: The Mann-Kendall test identified:

- **Positive Precipitation Trends:** Concentrated in northern high-altitude regions, potentially linked to orographic effects.
- **Negative Trends:** Notable in the southern plains, suggesting reduced monsoonal activity.
- The **p-value** indicates the probability that the observed trend is due to random chance. A **small p-value** suggests a statistically significant trend (generally people take <0.05).

Quantified Slopes: Sen's Slope Estimator calculated trends in mm/year. The results:

- The **Sen's Slope** is the **median** of all individual slopes. It highlights localized hotspots of climatic variability.
- Provided quantitative metrics for comparative spatial analysis.



Grid_Point	Trend	p-value	Slope	Latitude	Longitude	Color
23.5,92.5	decreasing	0.019068322	-0.0014849	23.5	92.5	red
23.5,93.5	decreasing	0.007701657	-0.0014669	23.5	93.5	red
24.5,91.5	decreasing	0.044331742	-0.001142	24.5	91.5	red
24.5,92.5	decreasing	0.029620323	-0.0011298	24.5	92.5	red
24.5,93.5	no trend	0.156908476	-0.000595	24.5	93.5	gray
24.5,94.5	no trend	0.089383574	-0.0007008	24.5	94.5	gray
25.5,90.5	no trend	0.070305407	-0.0005442	25.5	90.5	gray
25.5,91.5	no trend	0.782043272	-0.0001052	25.5	91.5	gray
25.5,92.5	no trend	0.092172361	-0.0007319	25.5	92.5	gray
25.5,93.5	no trend	0.15014508	-0.0005811	25.5	93.5	gray
25.5,94.5	decreasing	0.04595954	-0.0008142	25.5	94.5	red
26.5,90.5	no trend	0.273272975	-0.0002288	26.5	90.5	gray
26.5,91.5	no trend	0.599594695	-0.0002205	26.5	91.5	gray
26.5,92.5	no trend	0.244196284	-0.0004899	26.5	92.5	gray
26.5,93.5	no trend	0.297870496	-0.0004304	26.5	93.5	gray
26.5,94.5	decreasing	0.012857616	-0.0014205	26.5	94.5	red
26.5,95.5	decreasing	0	-0.0117546	26.5	95.5	red
27.5,90.5	no trend	0.195684802	-0.000427	27.5	90.5	gray
27.5,91.5	no trend	0.388687536	-0.0003481	27.5	91.5	gray
27.5,92.5	no trend	0.145698318	-0.0006255	27.5	92.5	gray
27.5,93.5	decreasing	0.013833266	-0.001143	27.5	93.5	red
27.5,94.5	decreasing	1.55E-06	-0.0051951	27.5	94.5	red
27.5,95.5	decreasing	1.83E-09	-0.0060612	27.5	95.5	red
27.5,96.5	decreasing	9.45E-10	-0.0062149	27.5	96.5	red
28.5,91.5	no trend	0.791445751	-9.17E-05	28.5	91.5	gray
28.5,92.5	no trend	0.652439481	-0.0001725	28.5	92.5	gray
28.5,93.5	decreasing	2.27E-07	-0.0049766	28.5	93.5	red
28.5,94.5	decreasing	1.85E-07	-0.0055993	28.5	94.5	red
28.5,95.5	decreasing	2.89E-09	-0.0058803	28.5	95.5	red

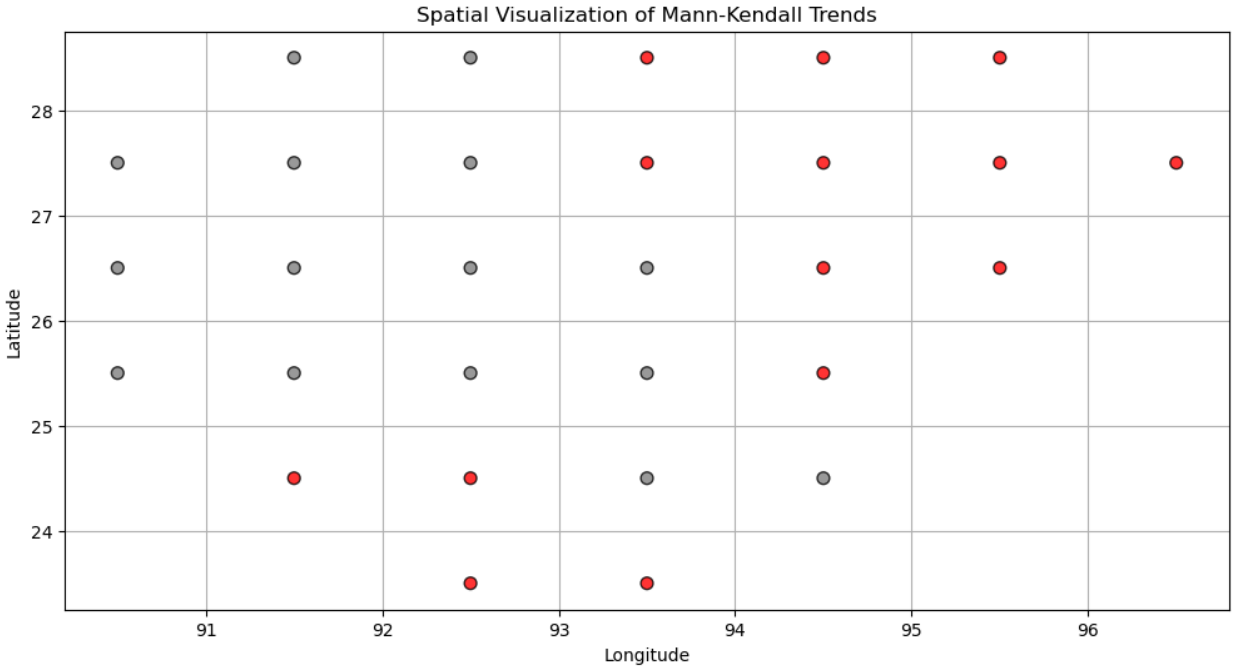


Fig. 3 Trend analysis

3.3 Temporal Analysis

3.3.1 Monthly Aggregation Analysis

- **Precipitation Peaks:** Observed during the monsoon season (June–September), consistent with the region's dependence on monsoonal rainfall that can be seen in fig. 4.
- **Dry Spells:** Winter months (December–February) showed negligible precipitation, aligning with historical drought patterns.

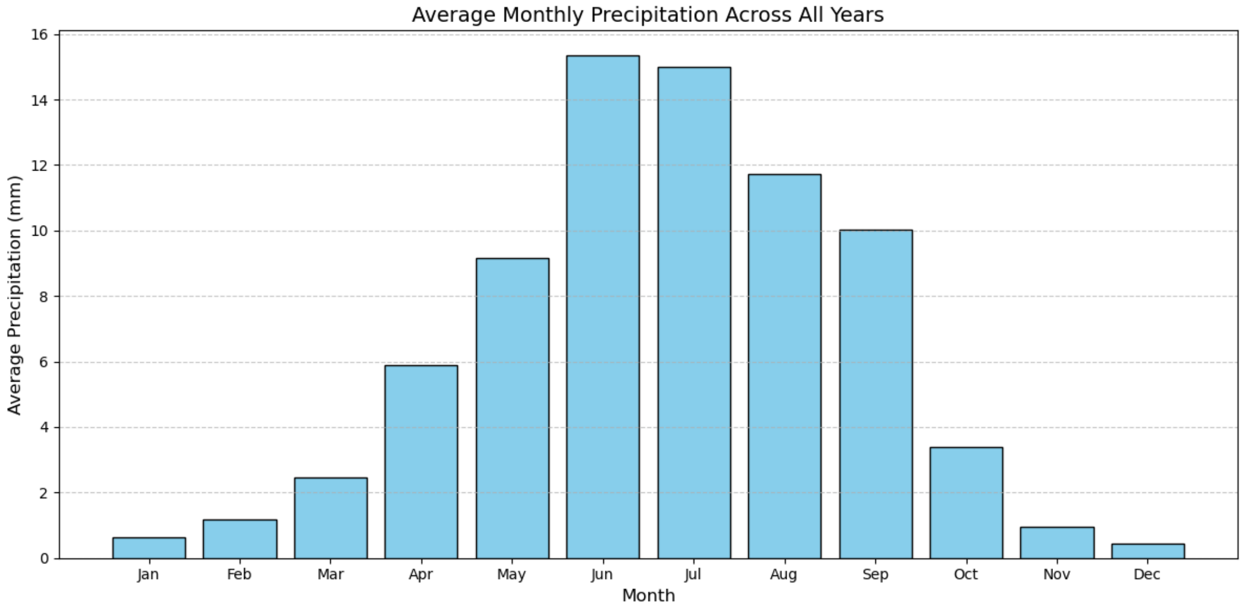


Fig. 4 Average monthly precipitation across all years

3.3.2 Seasonal Aggregation Analysis

- **Monsoonal Dominance**: Seasonal analysis highlighted the dominance of monsoon rainfall (June–September) across the region, contributing over 70% of the annual precipitation. This is clearly represented in the seasonal precipitation plot in Fig. 5, where the spatial intensity of rainfall aligns with Northeast India's dependence on monsoonal patterns.
- **Post-Monsoon Transition**: The post-monsoon period (October–November) exhibited moderate precipitation levels, reflecting the gradual withdrawal of the monsoon and localized rainfall events.

- **Winter Dryness**: Minimal precipitation was observed during winter months (December–February), indicating prolonged dry spells typical of the region's seasonal drought history. This is corroborated by the temporal patterns visualized in Fig. 4.

- **Pre-Monsoon Activity**: Early pre-monsoon months (March–May) showed rising precipitation levels, likely driven by convectional activity and localized thunderstorms.

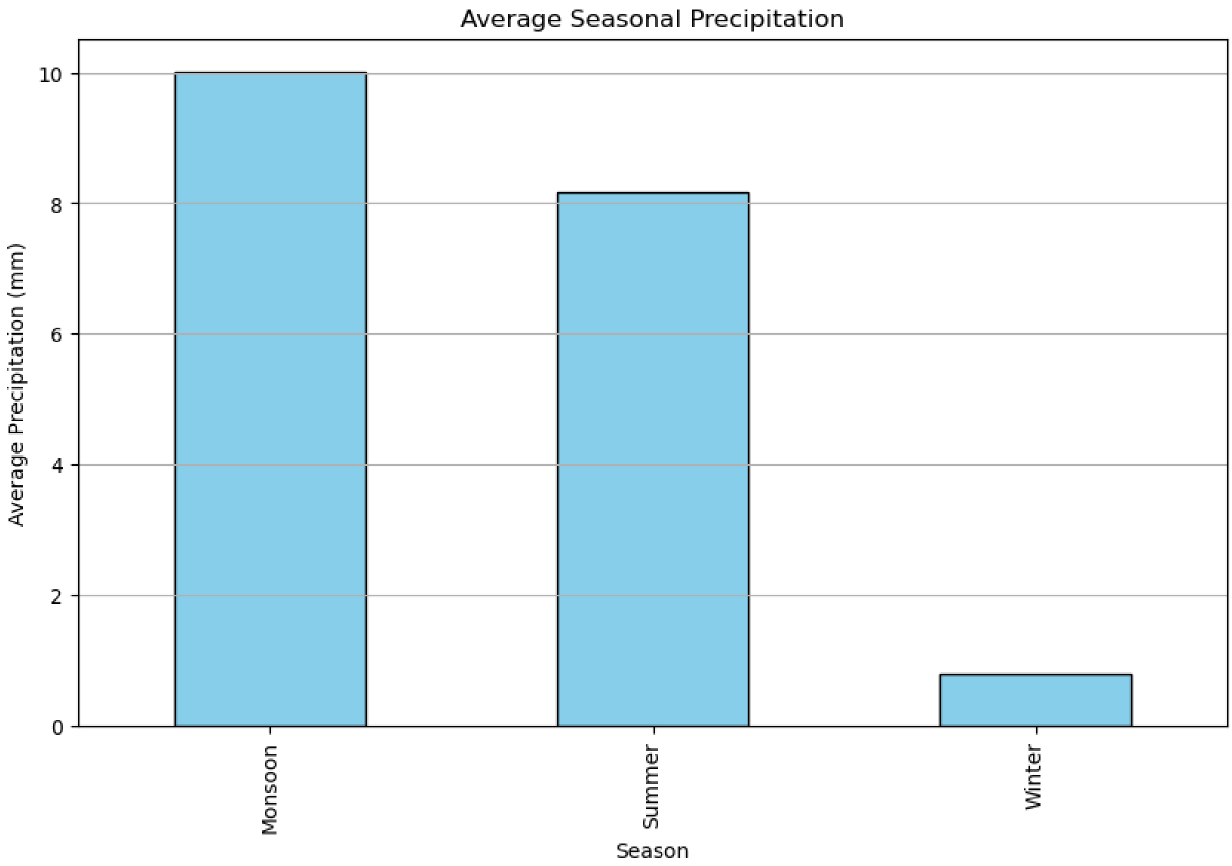


Fig. 5 Average seasonal precipitation

3.3.3 Time Series Analysis at Grid Level

The temporal distribution for a representative grid location (e.g., 23.5, 92.5) is shown in Fig. 6, illustrating clear seasonal cycles:

- **Peak Activity**: Noticeable peaks during monsoonal months with significant rainfall accumulation.

- **Periodic Variations**: Clear periodic oscillations reflecting seasonal climatic variability.

- **Low Noise**: Consistent patterns validate the preprocessing steps in removing outliers and ensuring data integrity.

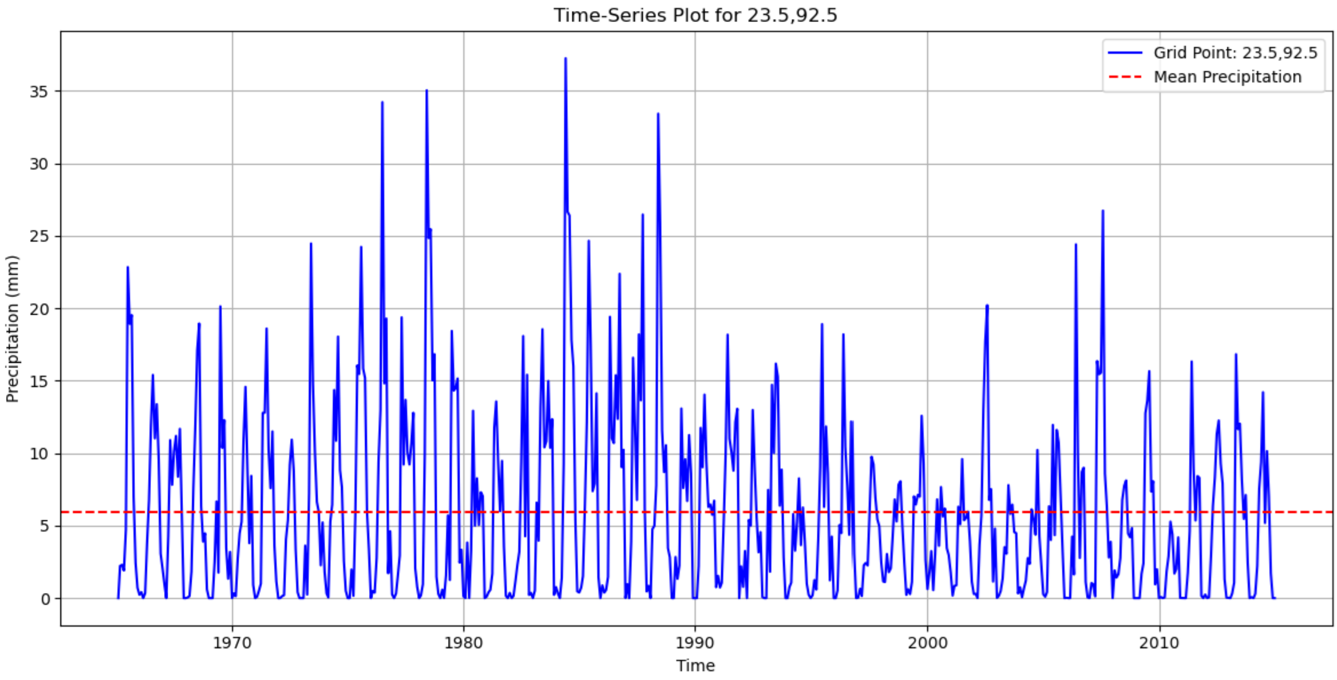


Fig. 6 Sample time series plot of latitude:23.5, longitude:92.5

3.4 Spatial Distribution

Spatial analysis of precipitation and temperature data provided critical insights into the geographic variability of climatic patterns across Northeast India. This region's diverse topography influences significant contrasts in precipitation intensity and distribution, which were effectively captured in the analysis.

3.4.1 High Precipitation Zones

Regions such as the northeastern highlands, prominently including Cherrapunji and nearby areas, emerged as hotspots for extreme precipitation. These regions are subject to:

- **Orographic Effects:** The high-altitude terrain significantly enhances rainfall as moist monsoonal winds ascend, cool, and condense over these elevations.

- **Annual Rainfall Extremes:** These areas often record some of the world's highest annual precipitation levels, underscoring their critical role in regional hydrological cycles.

- **Localized Rainfall Peaks:** High spatial variability was observed within these zones, indicative of micro-climatic influences driven by specific topographic and atmospheric interactions.

3.4.2 Dry Zones

Contrasting the high precipitation areas, the lower valleys and plains exhibited significantly reduced rainfall levels. This spatial pattern aligns with:

- **Rain-Shadow Effects:** Valleys situated leeward of the highlands receive diminished rainfall due to the obstruction of moist monsoonal winds by intervening topographic barriers.
- **Monsoonal Dependency:** These regions show a pronounced dependence on monsoonal rains, with reduced precipitation during non-monsoon periods.
- **Agricultural Implications:** The scarcity of rainfall in these dry zones poses challenges for rain-fed agriculture and water resource management, emphasizing the need for adaptive planning.

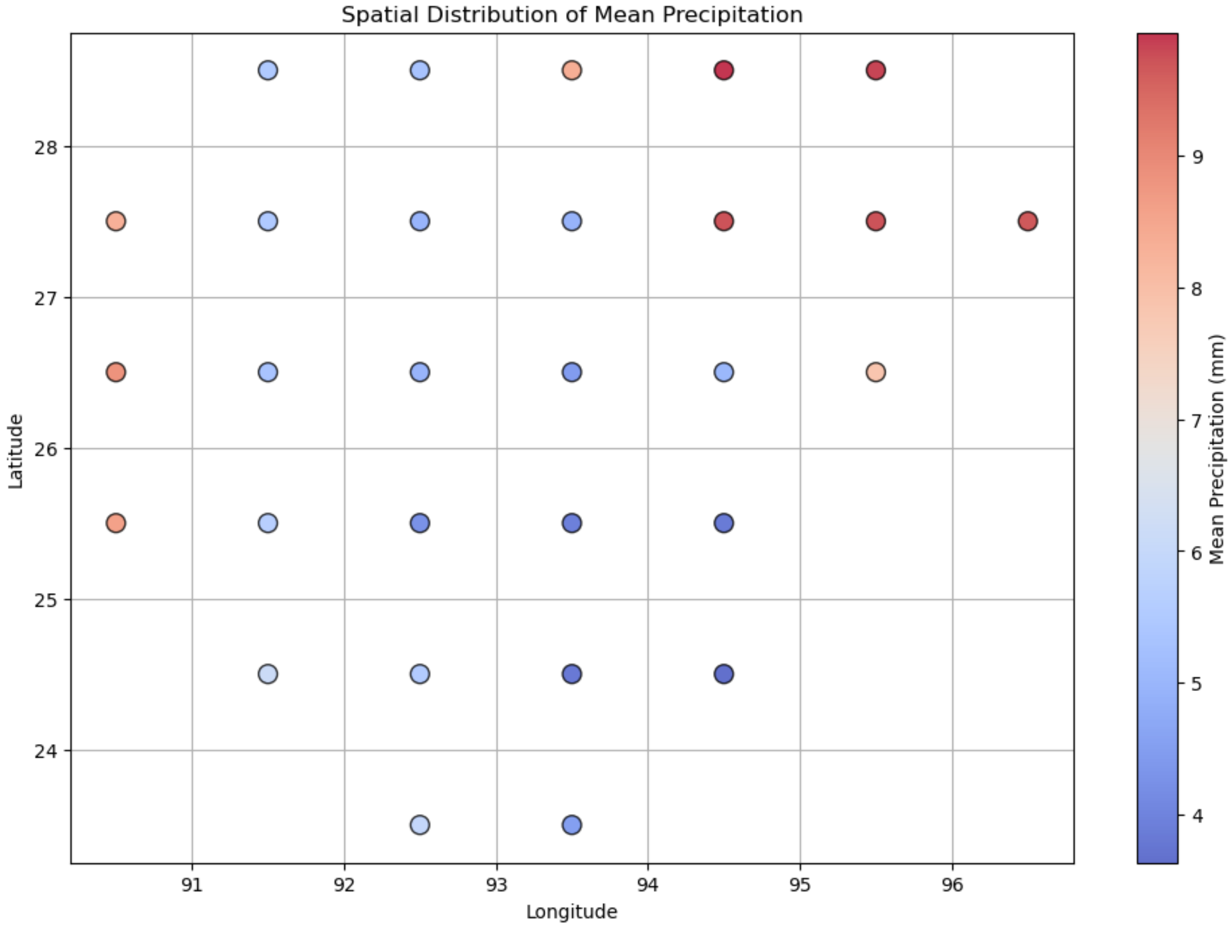


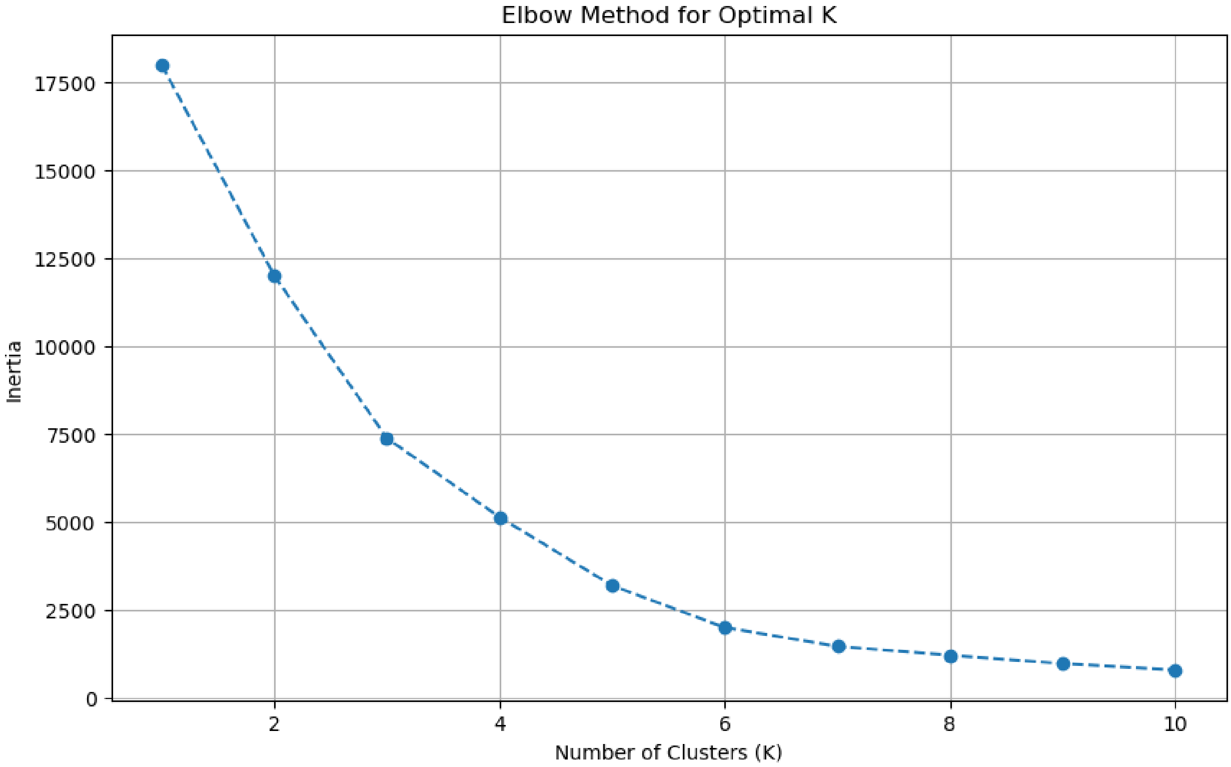
Fig. 7 Spatial distribution of mean precipitation

3.5. Clustering Analysis

3.5.1 K-Means Clustering

Performance Analysis:

- K-Means effectively segmented structured regions, identifying six distinct clusters based on precipitation and temperature distributions.
- To determine the optimal number of clusters (k), the elbow method and silhouette method was employed as represented in Fig. 8, which evaluates the cohesion and separation of clusters to identify the most suitable value for k, ensuring well-defined and meaningful clusters.
- We used optimal_K =6 based on silhouette method result.



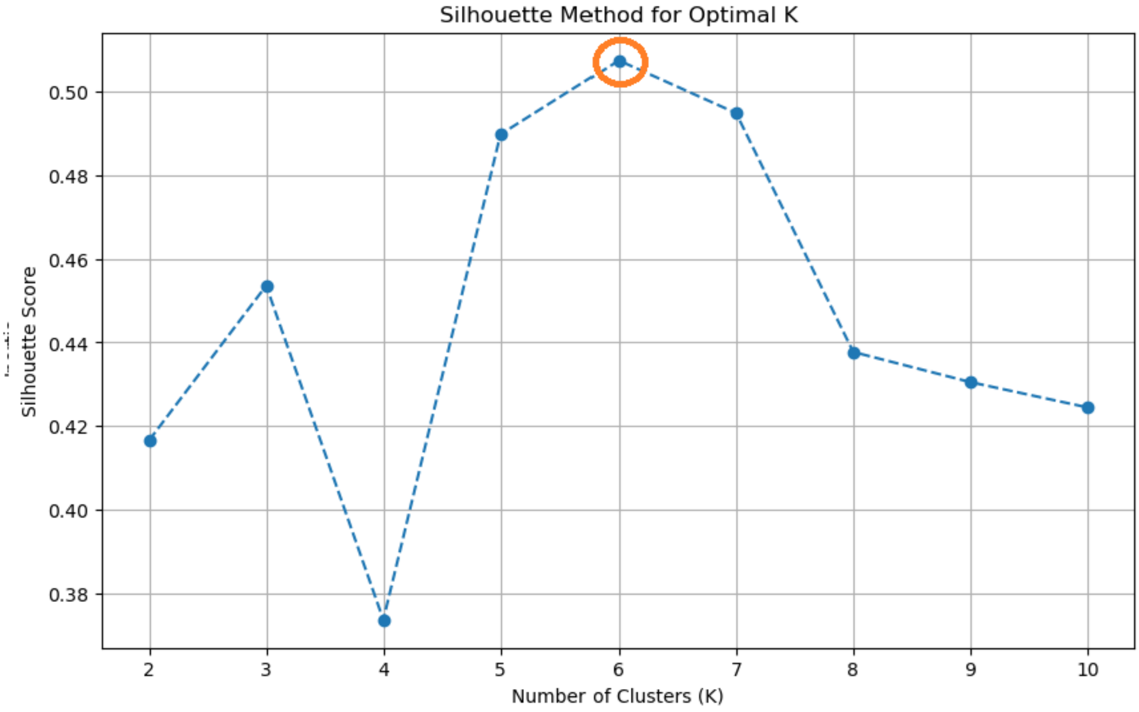


Fig. 8 Methods for finding optimal K

Cluster Insights:

- Regions with homogeneous precipitation patterns, such as the Brahmaputra Valley, formed uniform clusters.
- High-elevation areas, like Meghalaya, emerged as unique clusters due to distinct rainfall patterns.

Advantages:

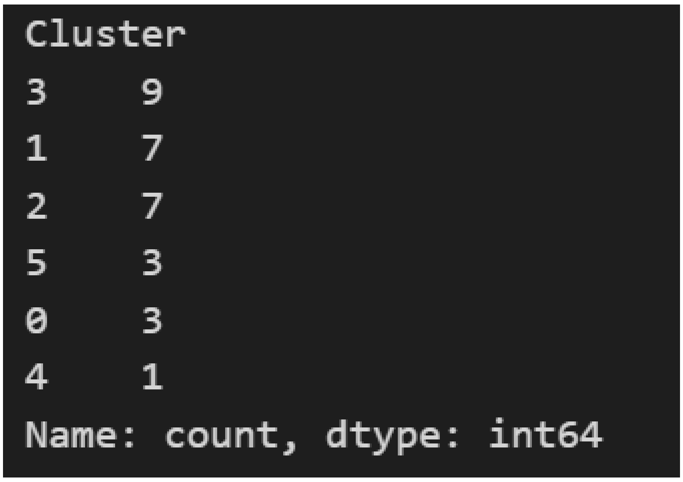
- Efficient for structured datasets with low noise.
- Computational simplicity allowed scalability for large datasets.

Limitation:

- Failed to capture irregular or non-linear patterns, as evidenced by the merging of heterogeneous climatic zones in some regions.

Clusters:

We obtained 6 clusters and namely cluster-3 had the highest number of grid-point in it as shown in Fig. 9, it also visually shows the representation of clusters obtained from K-mean clustering algorithm on precipitation data.



```
Cluster
3    9
1    7
2    7
5    3
0    3
4    1
Name: count, dtype: int64
```

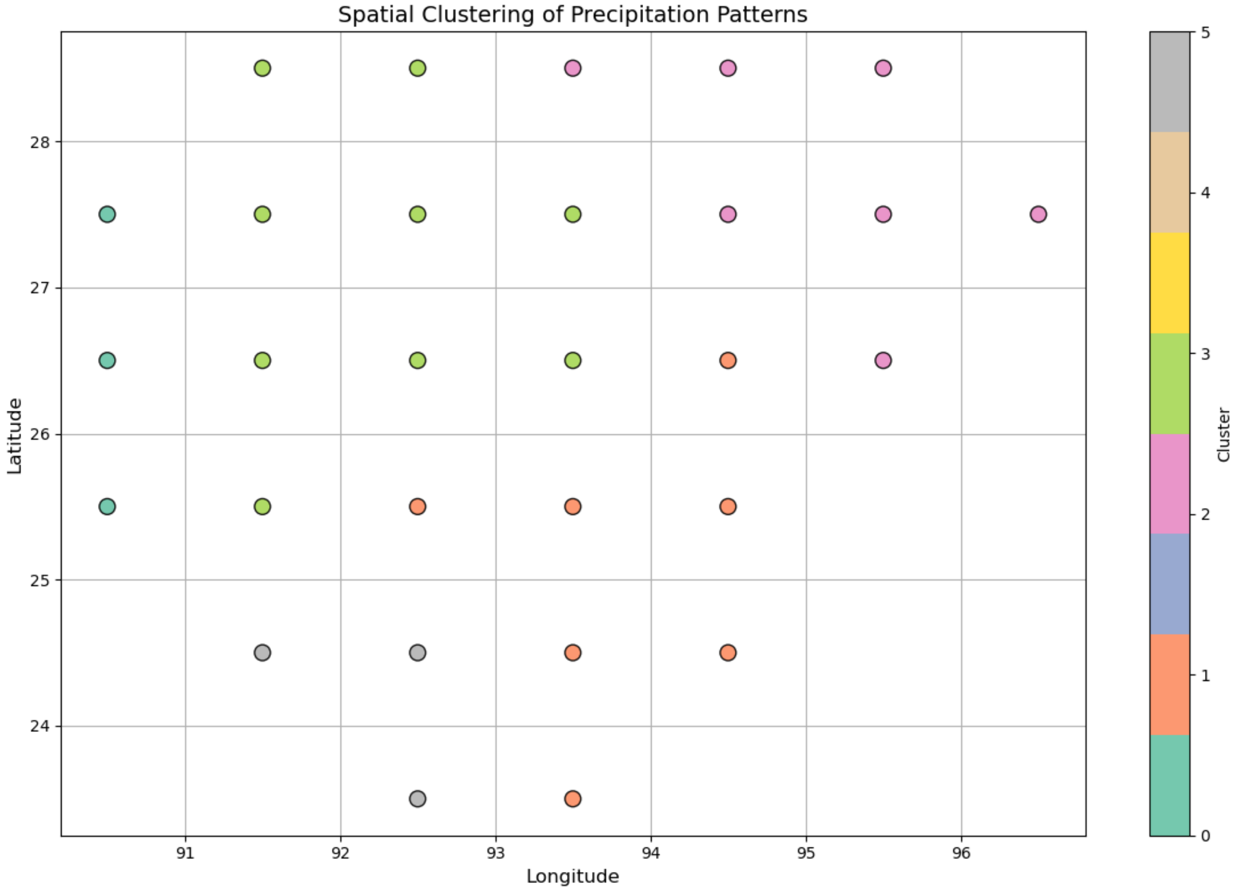


Fig. 9 K-means clustering

3.5.2 DBSCAN Clustering

Robustness to Noise:

- DBSCAN proved effective in isolating anomalous grid points as noise, particularly those affected by extreme climatic events like flash floods.
- **Optimal Parameters:**
  - Eps=105.0 and Min_samples=7, determined using a k-distance graph (Fig. 10).

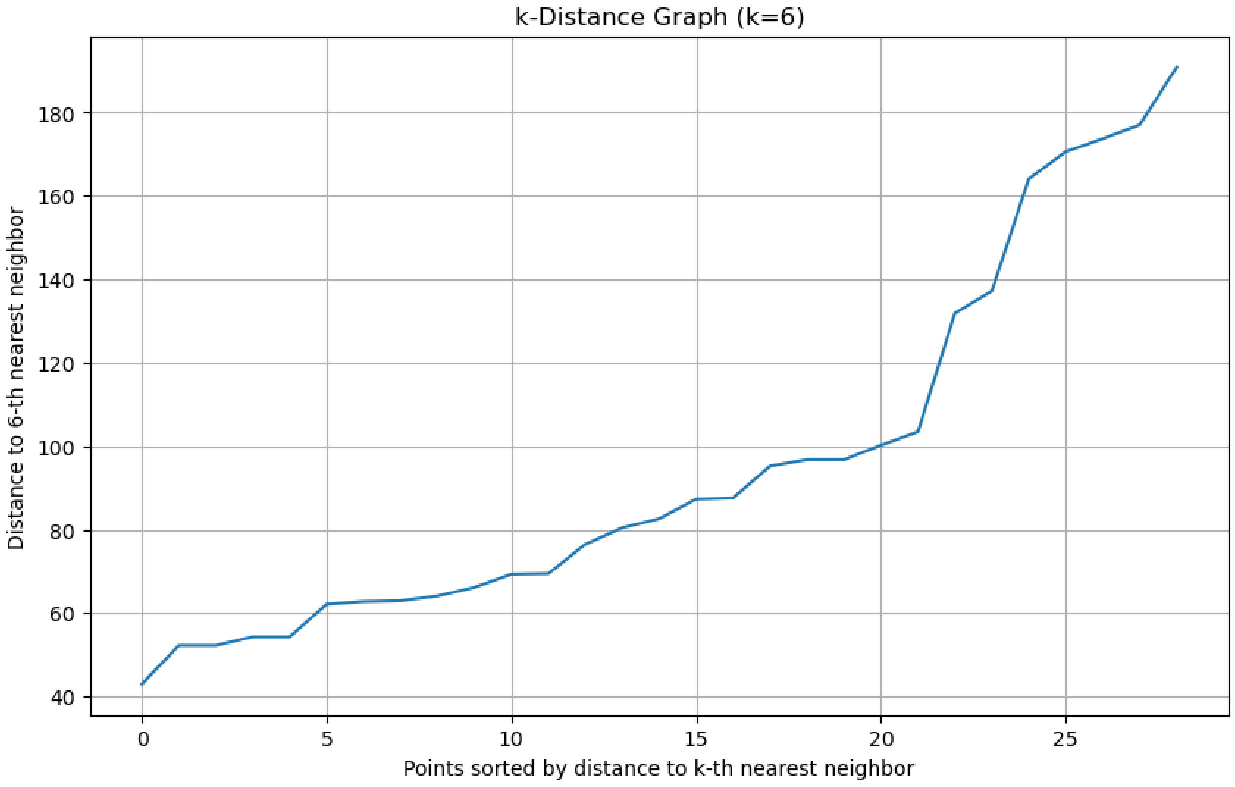


Fig. 10 Plot to find optimal parameters of DBSCAN

Key Findings:

- **Irregular Clusters:** Highlighted unique patterns in highly variable regions, such as areas impacted by shifting monsoon patterns.
- **Anomalies as Noise:** Noise points captured rare events, aiding in disaster risk mapping.

Advantages:

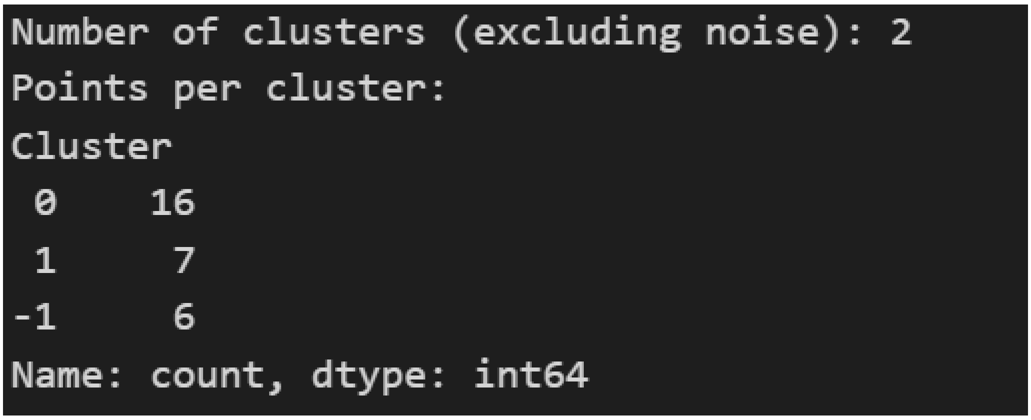
- Adaptable to irregular spatial and temporal patterns.
- Resilient to outliers.

Limitations:

- Parameter sensitivity required extensive tuning, particularly in densely sampled regions.

Clusters:

- We obtained 3 clusters and namely cluster-0 had the highest number of grid-point in it as shown in Fig. 11, it also visually shows the representation of clusters obtained from DBSCAN clustering algorithm on precipitation data.



```
Number of clusters (excluding noise): 2
Points per cluster:
Cluster
 0    16
 1     7
-1     6
Name: count, dtype: int64
```

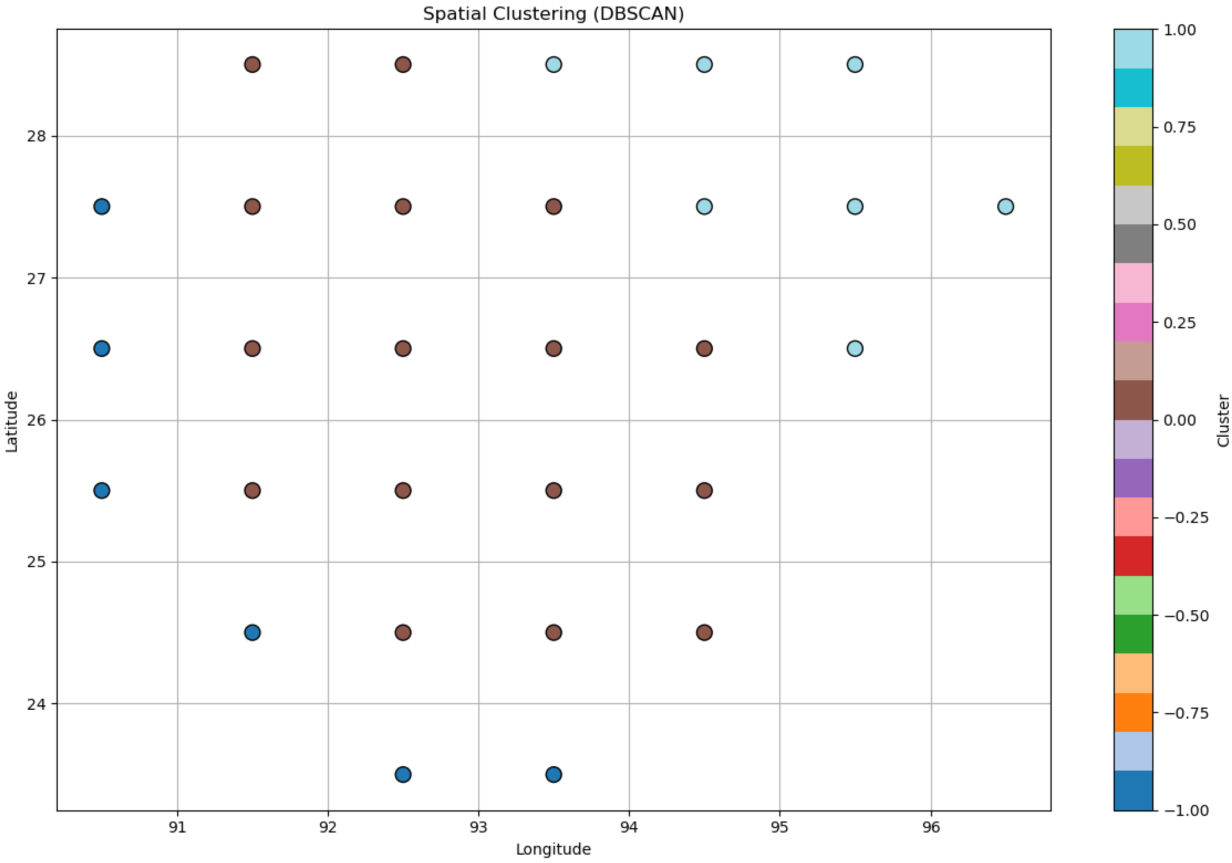


Fig. 11 DBSCAN clustering

.

3.5.3 Hierarchical Clustering

Uncovering Nested Relationships:

- Hierarchical clustering provided insights into nested climatic behaviors, with clusters revealing multi-scale patterns.

Cluster Characteristics:

- Larger regions were hierarchically subdivided, offering granular insights into sub-regional climate dynamics.
- Dendrogram analysis was performed using Ward's method (Fig. 12) revealed clear parent-child relationships among climatic zones.

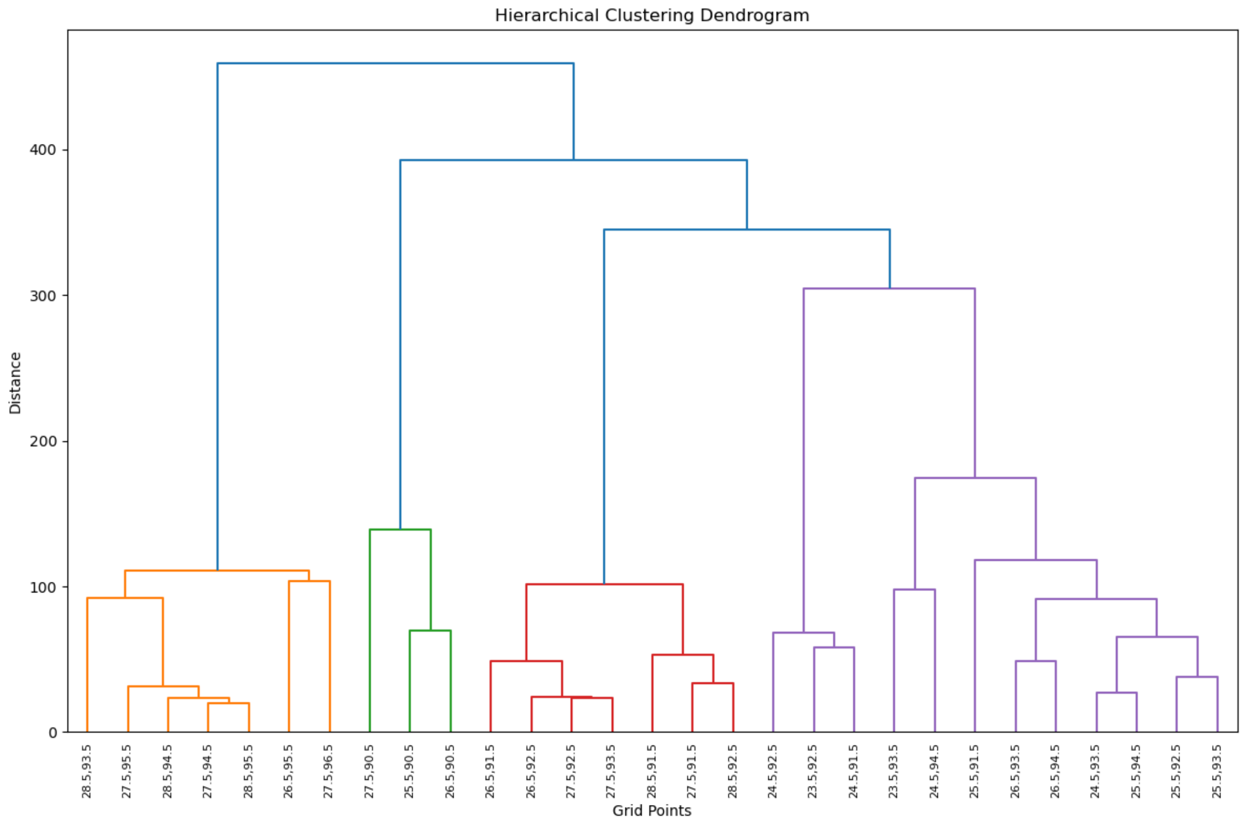


Fig. 12 Dendrogram analysis

Advantages:

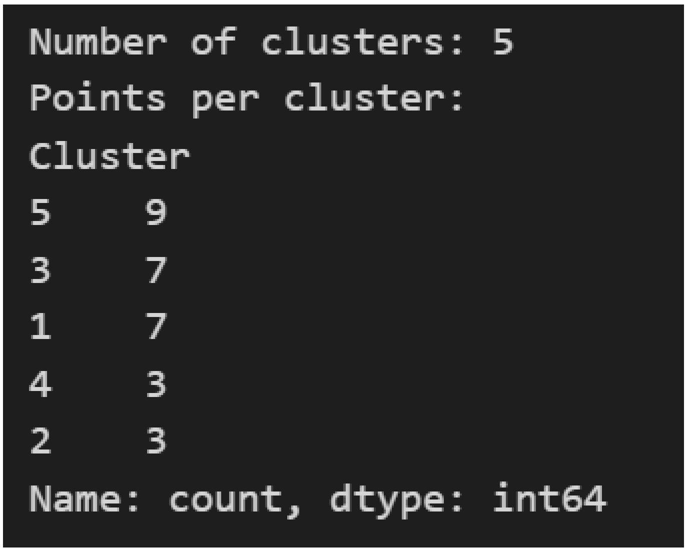
- Ideal for exploratory analyses requiring a hierarchical understanding of climatic systems.
- Captured nested relationships that other methods missed.

Limitations:

- Computational demands were high, especially for larger datasets.

Clusters:

- We obtained 5 clusters and namely cluster-5 had the highest number of grid-point in it as shown in Fig. 13, it also visually shows the representation of clusters obtained from Hierarchical clustering algorithm on precipitation data.



```
Number of clusters: 5
Points per cluster:
Cluster
5    9
3    7
1    7
4    3
2    3
Name: count, dtype: int64
```

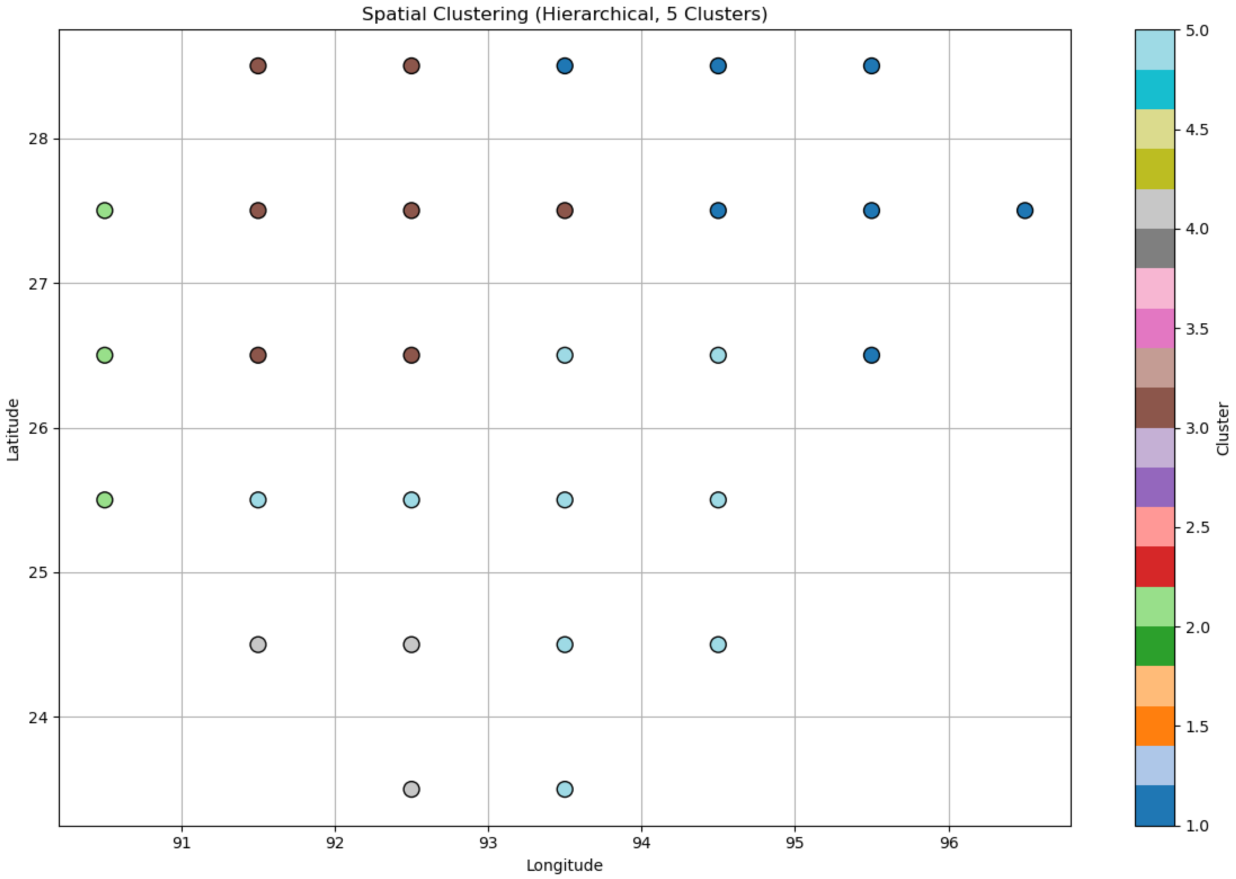


Fig. 13 Hierarchical Clustering

Chapter 4

Results and Discussion

4.1 Data Preprocessing and Trend Analysis

4.1.1 Handling Missing Values

Addressing data gaps in climatic datasets was critical for maintaining the integrity of the analysis. Missing values, arising from sensor malfunctions and data transmission issues, were treated through temporal interpolation and grid-wise mean imputation. Temporal interpolation preserved time-series trends, while grid-wise imputation minimized spatial distortions, ensuring reliable input for clustering methods.

4.1.2 Temporal Aggregation

The transformation of daily data into monthly and seasonal aggregates highlighted key climatic trends. Monthly aggregation revealed variability, particularly monsoonal dominance, whereas seasonal aggregation segmented data into winter, summer, and monsoon periods, facilitating intra-annual analyses.

4.1.3 Standardization

Z-score normalization mitigated biases caused by extreme values, ensuring that precipitation intensity and temperature ranges were equally weighted across all grid points. This step was pivotal for enhancing the robustness of clustering methods.

4.2 Distribution Analysis

4.2.1 Mann-Kendall Test and Sen’s Slope Estimator

The Mann-Kendall test identified significant trends:

- **Positive Trends**: Concentrated in high-altitude northern regions, likely driven by orographic effects.
- **Negative Trends**: Observed in southern plains, suggesting reduced monsoonal activity.

Sen’s Slope Estimator quantified these trends, offering a metric for comparing spatial climatic variations. The slopes provided critical insights into regions experiencing intensified precipitation or prolonged dry spells.

4.2.2 Temporal Distribution

Temporal analyses highlighted key patterns:

- **Monsoonal Dominance**: June–September accounted for over 70% of annual precipitation, underscoring the region’s reliance on monsoonal rainfall.
- **Winter Droughts**: Negligible rainfall during December–February revealed consistent seasonal dry spells.
- **Time-Series Patterns**: A representative grid point (23.5°N, 92.5°E) demonstrated clear periodic oscillations, validating preprocessing and reflecting seasonal climatic cycles.

4.2.3 Spatial Distribution

High Precipitation Zones

High-altitude regions such as Cherrapunji and Mawsynram exhibited extreme precipitation, attributed to:

- **Orographic Effects**: Moist monsoonal winds condensing over high-altitude terrains.
- **Localized Peaks**: Micro-climatic interactions enhanced rainfall variability within these hotspots.

Dry Zones

Contrasting high rainfall zones, valleys and plains exhibited reduced precipitation, influenced by:

- **Rain-Shadow Effects**: Valleys shielded from monsoonal winds by highlands.
- **Monsoonal Dependence**: These regions relied heavily on monsoon rains, highlighting vulnerability during non-monsoon periods.

4.3 Clustering Analysis

4.3.1 K-Means Clustering

Using the silhouette method, six optimal clusters were identified, highlighting:

- **Homogeneous Regions**: Structured zones like the Brahmaputra Valley formed uniform clusters.
- **Distinct High-Elevation Areas**: Meghalaya emerged as a unique cluster due to significant rainfall disparities.

Advantages included computational efficiency and scalability for large datasets, while limitations included challenges in capturing irregular patterns. Fig. 9 visually illustrates the cluster representation, with Cluster 3 encompassing the highest number of grid points.

4.3.2 DBSCAN Clustering

DBSCAN, using optimized parameters (eps=105.0, min_samples=7), identified:

- **Irregular Clusters**: Captured unique patterns in highly variable regions.
- **Anomalous Events**: Isolated outliers representing rare climatic phenomena such as flash floods.

This method excelled in handling noise and detecting non-linear patterns but required extensive parameter tuning. Fig. 11 displays DBSCAN clusters, with Cluster 0 encompassing most grid points.

4.3.3 Hierarchical Clustering

Hierarchical clustering, performed using Ward’s method, uncovered:

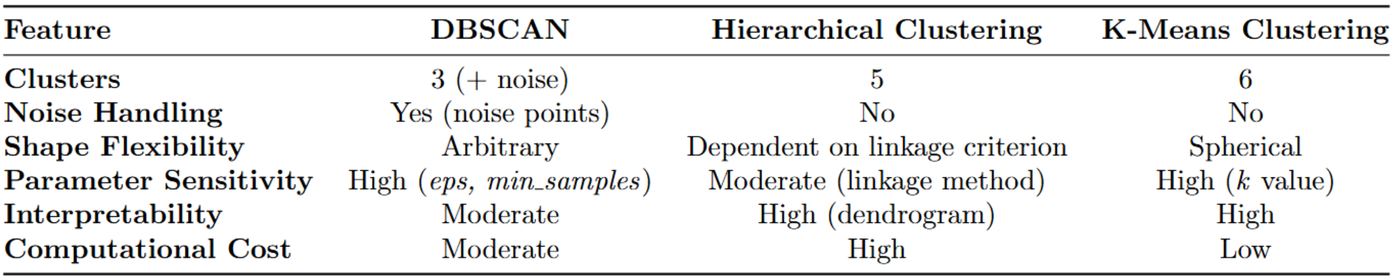
- **Nested Relationships**: Highlighted multi-scale climatic behaviors.
- **Granular Subdivisions**: Sub-regional dynamics provided detailed insights into spatial variability.

The dendrogram (Fig. 12) revealed parent-child hierarchies, while Cluster 5 contained the highest number of grid points, as shown in Fig. 13. Although computationally intensive, this method captured patterns missed by other algorithms.

4.4 Comparative Analysis of Clustering Methods

The clustering techniques provided complementary insights (Fig. 14):

- **K-Means**: Effective for structured datasets; computationally efficient but less adaptive to noise.
- **DBSCAN**: Robust for noise handling and non-linear patterns; parameter sensitive.
- **Hierarchical Clustering**: Ideal for exploratory analysis and nested patterns; computationally demanding.



Feature	DBSCAN	Hierarchical Clustering	K-Means Clustering
Clusters	3 (+ noise)	5	6
Noise Handling	Yes (noise points)	No	No
Shape Flexibility	Arbitrary	Dependent on linkage criterion	Spherical
Parameter Sensitivity	High (*eps*, *min_samples*)	Moderate (linkage method)	High (k value)
Interpretability	Moderate	High (dendrogram)	High
Computational Cost	Moderate	High	Low

Fig. 14 Key Comparisons Between Clustering Methods

4.5 Key Findings

- **Precipitation Hotspots**: High rainfall zones identified critical areas for hydrological planning.
- **Dry Zones**: Highlighted the need for adaptive strategies in water-scarce regions.
- **Clustering Insights**: Demonstrated the utility of advanced clustering methods in delineating spatial-temporal climate patterns.

Chapter 5

Conclusion

This study systematically evaluates the application of advanced clustering methodologies for the spatio-temporal analysis of precipitation and temperature data in Northeast India. By leveraging K-Means, DBSCAN, and Hierarchical Clustering techniques, the research elucidates significant climatic patterns, including the delineation of precipitation hotspots, temperature gradients, and anomalous regions. Each clustering algorithm exhibited distinct technical advantages: K-Means demonstrated high efficiency in segmenting structured datasets, DBSCAN excelled in identifying irregular clusters and managing noise, and Hierarchical Clustering effectively captured nested and multi-scale climatic relationships. The outcomes validate the applicability of clustering methods as robust analytical tools for climate data interpretation and strategic planning.

5.2 Future Work

In the next phase of this research, similar analyses will be extended to temperature datasets, with a focus on incorporating supplementary data sources, such as land use patterns and atmospheric circulation models. This integration aims to enhance clustering accuracy and reliability. Additionally, the findings will be validated through empirical evaluations under real-world climatic conditions, ensuring their practical applicability for regional climate planning and mitigation strategies.

References

1. Goyal, M. K., Shivam, G., and Sarma, A. K. (2019), “Spatial Homogeneity of Extreme Precipitation Indices Using Fuzzy Clustering over Northeast India.” *Environmental Earth Sciences*, 78(2), 45-55.
2. Salehnia, N., Salehnia, N., Ansari, H., Kolsoumi, S., and Bannayan, M. (2019), “Climate Data Clustering Effects on Arid and Semi-Arid Rainfed Wheat Yield: A Comparison of Artificial Intelligence and K-Means Approaches.” *Field Crops Research*, 234, 12-23.
3. Yang, P., Xia, J., Zhang, Y., Han, J., and Wu, X. (2017), “Quantile Regression and Clustering Analysis of Standardized Precipitation Index in the Tarim River Basin, Xinjiang, China.” *Hydrological Processes*, 31(10), 2069-2084.
4. Yu, Y., Shao, Q., and Lin, Z. (2020), “Regionalization Study of Maximum Daily Temperature Based on Grid Data by an Objective Hybrid Clustering Approach.” *Journal of Hydrology*, 586, 124-137.
5. Jin, H., Chen, X., Wu, P., Song, C., and Xia, W. (2019), “Evaluation of Spatial-Temporal Distribution of Precipitation in Mainland China by Statistic and Clustering Methods.” *Atmospheric Research*, 217, 23-38.
6. Zerouali, B., Chettih, M., Abda, Z., Mesbah, M., Santos, C. A. G., and Neto, R. M. B. (2022), “A New Regionalization of Rainfall Patterns Based on Wavelet Transform Information and Hierarchical Cluster Analysis in Northeastern Algeria.” *Springer-Verlag*, 45, 11-30.
7. Lin, W., Liu, Y., Li, N., Wang, J., Zhang, N., Wang, Y., Wang, M., Ren, H., and Li, M. (2020), “Study of the Spatiotemporal Distribution Characteristics of Rainfall Using Hybrid Dimensionality Reduction-Clustering Model: A Case Study of Kunming City, China.” *Science of the Total Environment*, 706, 135-145.
8. Singh, A., Thakur, S., and Adhikary, N. C. (2021), “Analysis of Spatial and Temporal Rainfall Characteristics of the North East Region of India.” *Journal of Atmospheric and Climate Science*, 16(5), 97-112.
9. Subash, N., Sikka, A. K., and Ram Mohan, H. S. (2010), “An Investigation into Observational Characteristics of Rainfall and Temperature in Central Northeast India—A Historical Perspective 1889–2008.” *International Journal of Climatology*, 30(7), 1107-1117.

10. Kumari, M., Singh, C. K., and Basistha, A. (2016), "Clustering Data and Incorporating Topographical Variables for Improving Spatial Interpolation of Rainfall in Mountainous Region." *Hydrology Research*, 47(3), 598-612.

11. Ramos, M. C. (2001), "Divisive and Hierarchical Clustering Techniques to Analyse Variability of Rainfall Distribution Patterns in a Mediterranean Region." *Atmospheric Research*, 57(3), 307-315.

12. Golian, S., Saghafian, B., Sheshangosht, S., and Ghalkhani, H. (2010), "Comparison of Classification and Clustering Methods in Spatial Rainfall Pattern Recognition at Northern Iran." *Meteorological Applications*, 17(4), 567-576.

13. Dash, S. K., Sharma, N., Pattnayak, K. C., Gao, X. J., Shi, Y. (2012), "Temperature and Precipitation Changes in the North-East India and Their Future Projections." *Global and Planetary Change*, 98, 31-44.

14. Groisman, P. Y., Karl, T. R., Easterling, D. R., Knight, R. W., Jamason, P. F., et al. (1999), "Changes in the Probability of Heavy Precipitation: Important Indicators of Climatic Change." *Climatic Change*, 42(1), 243-283.

15. Manton, M. J., Della-Marta, P. M., Haylock, M. R., Hennessy, K. J., Nicholls, N., et al. (2001), "Trends in Extreme Daily Rainfall and Temperature in Southeast Asia and the South Pacific: 1961–1998." *International Journal of Climatology*, 21(3), 269-284.

16. Basistha, A., Arya, D. S., and Goel, N. K. (2008), "Spatial Distribution of Rainfall in Indian Himalayas – A Case Study of Uttarakhand Region." *Water Resources Management*, 22(8), 1325-1346.

17. Wackernagel, H. (2003), "Multivariate Geostatistics." *Springer*, Berlin Heidelberg, 2nd ed., 343-349.

18. Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. E. (2011), "Probability and Statistics for Engineers and Scientists." *9th Edition*, Pearson Education, Boston, USA.

19. Rao, A. R., and Srinivas, V. V. (2008), "Regionalization of Watersheds: An Approach Based on Cluster Analysis." *Water Science and Technology Library*, 58, Springer, Dordrecht, Netherlands, pp. 103-119.

20. Mellak, S., and Souag-Gamane, D. (2020), "Spatio-Temporal Analysis of Maximum Drought Severity Using Copulas in Northern Algeria." *Journal of Water and Climate Change*, 11(1), 49-59.

21. Yim, O., and Ramdeen, K. T. (2015), "Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data." *Quantitative Methods in Psychology*, 11(1), 8-21.