

Article

Study of the Spatiotemporal Distribution Characteristics of Rainfall Using Hybrid Dimensionality Reduction-Clustering Model: A Case Study of Kunming City, China

Weijie Lin ^{1,2,3}, Yuanyuan Liu ^{1,2,3}, Na Li ^{1,2,3,*}, Jing Wang ^{1,2,3}, Nianqiang Zhang ^{1,2,3}, Yanyan Wang ^{1,2,3}, Mingyang Wang ^{1,2,3} , Hancheng Ren ^{1,2,3} and Min Li ^{1,2,3}

- ¹ State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
² China Institute of Water Resources and Hydropower Research, Beijing 100038, China
³ Research Center on Flood and Drought Disaster Prevention and Reduction of the Ministry of Water Resources, Beijing 100038, China
* Correspondence: lina@iwhr.com

Abstract: In recent years, the frequency and intensity of global extreme weather events have gradually increased, leading to significant changes in urban rainfall patterns. The uneven distribution of rainfall has caused varying degrees of water security issues in different regions. Accurately grasping the spatiotemporal distribution patterns of rainfall is crucial for understanding the hydrological cycle and predicting the availability of water resources. This study collected rainfall data every five minutes from 62 rain gauge stations in the main urban area of Kunming City from 2019 to 2021, constructing an unsupervised hybrid dimensionality reduction-clustering (HDRC) model. The model employs the Locally Linear Embedding (LLE) algorithm from manifold learning for dimensionality reduction of the data samples and uses the dynamic clustering K-Means algorithm for cluster analysis. The results show that the model categorizes the rainfall in the Kunming area into three types: The first type has its rainfall center distributed on the north shore of Dian Lake and the southern part of Kunming's main urban area, with spatial dynamics showing the rainfall distribution gradually developing from the Dian Lake water body towards the land. The second type's rainfall center is located in the northern mountainous area of Kunming, with a smaller spatial dynamic change trend. The water vapor has a relatively fixed and concentrated rainfall center due to the orographic uplift effect of the mountains. The third type's rainfall center is located in the main urban area of Kunming, with this type of rainfall showing smaller variations in all indicators, mainly occurring in May and September when the temperature is lower, related to the urban heat island effect. This research provides a general workflow for spatial rainfall classification, capable of mining the spatiotemporal distribution patterns of regional rainfall based on extensive data and generating typical samples of rainfall types.



Citation: Lin, W.; Liu, Y.; Li, N.; Wang, J.; Zhang, N.; Wang, Y.; Wang, M.; Ren, H.; Li, M. Study of the Spatiotemporal Distribution Characteristics of Rainfall Using Hybrid Dimensionality Reduction-Clustering Model: A Case Study of Kunming City, China. *Atmosphere* **2024**, *15*, 534. <https://doi.org/10.3390/atmos15050534>

Academic Editor: Jie Wu

Received: 17 March 2024

Revised: 19 April 2024

Accepted: 24 April 2024

Published: 26 April 2024

Keywords: spatiotemporal pattern of rainfall; LLE; K-Means; urban precipitation; machine learning; Kunming



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As global climate change intensifies, the frequency and intensity of extreme weather events are on an upward trend [1,2]. Moreover, due to the impact of human activities, urban areas are experiencing significant changes in rainfall patterns [3]. Water contributes to GDP and plays an important role in the economy in general [4]; precipitation variations have a heavy economic impact. These changes not only affect local agricultural production but also pose new challenges for urban water resource management and planning [5]. For instance, the uneven distribution of rainfall may lead to water scarcity in some regions, while others might face flooding disasters [6,7].

Precipitation is a key meteorological forcing factor in numerous geographical studies [8]. Accurately understanding the spatiotemporal distribution of rainfall data are crucial for comprehending hydrological cycles, predicting water resource availability, and devising effective flood prevention measures [2,9]. For example, from the perspective of flood risk, the spatiotemporal distribution of rainfall processes is one of the significant factors affecting flood behavior [10]. The evolution of hydrological models from lumped to distributed approaches highlights an essential feature: the ability to respond to changes in the spatiotemporal distribution of rainfall [11,12]. From a temporal perspective, the distribution of rainfall over time can significantly impact the runoff generation mechanisms of the underlying surface. Daily variability in conjunction with an evolving climate [13] plays an important role. The position of peak rainfall during precipitation events notably influences the magnitude of flooding hazards [14]. From a spatial perspective, the spatial distribution of rainfall influences the converge mechanisms of the underlying surface. Factors such as the slope, roughness, drainage capacity, and proximity to river channels of the rainfall's center on the ground projection show varying runoff characteristics in specific basins, leading to flood events with distinct spatial features [15]. Traditional rainfall analysis methods primarily rely on statistical techniques to assess rainfall patterns and their impacts on the environment and society [16]. Time series analysis, including moving averages and trend analysis (utilizing methods such as the Mann–Kendall trend test), is widely used to identify long-term trends and temporal correlations in rainfall data [17]. In spatial distribution analysis, Geographic Information System (GIS) technology and Kriging interpolation methods enable researchers to map the spatial distribution of rainfall and predict precipitation at unmeasured locations, revealing geographic characteristics of rainfall patterns [18]. Frequency analysis, such as extreme value analysis and recurrence interval analysis, assesses the frequency and intensity of extreme rainfall events, often employing probability distributions like Gumbel or Weibull for data analysis [19]. Furthermore, integrated approaches, including multivariate and regression analysis, combine various meteorological or geographical factors to analyze their collective impact on rainfall patterns [20]. Although these traditional methods have provided significant insights into understanding rainfall data over the past decades, they exhibit limitations in handling large datasets, identifying complex spatiotemporal dependencies, and analyzing extreme events [21]. Moreover, these methods often commence with statistical indicators of rainfall data and require further processing to reconstruct the processes of rainfall events with spatiotemporal distribution [22].

Consequently, with the advancement of computational technology, an increasing number of studies have begun to explore the use of advanced data analysis techniques to overcome these challenges, offering more in-depth analyses and comprehensive understandings. In recent years, the development of artificial intelligence, particularly unsupervised learning, has shown immense potential in analyzing complex meteorological datasets [23–26]. Unsupervised learning, a branch of machine learning, focuses on automatically discovering information and patterns in unlabeled data [27,28]. This is primarily achieved through two fundamental approaches: dimensionality reduction and clustering. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), aim to reduce the dimensions of data while preserving the most crucial information, thereby enhancing data processing efficiency and algorithm performance [29,30]. Clustering algorithms, including K-Means clustering and DBSCAN, reveal the intrinsic structure and patterns of data by grouping it into multiple categories with similar characteristics [31]. Unsupervised learning, by analyzing the characteristics and structure of data itself without explicit labels, offers a powerful tool for data mining and knowledge discovery [32,33]. Unsupervised learning methods are now widely used for classifying meteorological data and identifying extreme rainfall events. For instance, Liu Yuan Yuan et al. employed unsupervised learning algorithms to identify rainfall types in the Shenzhen area [34]. Henning Oppel and colleagues utilized a novel unsupervised learning algorithm to explore the patterns of rainfall timing distribution

and their correlation with flood type [35]. Andrew Mercer applied unsupervised learning to study the climate patterns of warm-season precipitation in the southeastern United States [36]. However, such studies rarely explore the rainfall patterns of a target region simultaneously from multiple aspects including the temporal distribution, spatial patterns, and statistical analysis of classification results.

Kunming, located in Yunnan Province, China, experiences a mild subtropical highland monsoon climate due to its geographical location and high altitude. The city enjoys a year-round mild climate with all four seasons resembling spring, and precipitation distribution is uneven, predominantly concentrated in the summer. Precipitation in Kunming is characterized by distinct seasonal patterns and diurnal variations. From a spatiotemporal perspective, precipitation begins to increase in May, peaks in September, and then declines rapidly. Peak rainfall typically occurs in the summer months, especially in July and August. Additionally, precipitation exhibits spatial variations, with mountainous regions surrounding the city receiving more rainfall than the urban plains due to topographical influences. Wang et al. selected precipitation data from Kunming's monitoring stations from 1998 to 2011 to study local extreme weather events [37]. Zhang et al. analyzed forty years of precipitation data up to 2020 to examine the impacts of urbanization on Kunming's rainfall patterns [38]. However, urbanization in this region has not ceased. According to our understanding, recent studies on the spatiotemporal distribution of urban precipitation in Kunming using recent data have been insufficient. Previous research has predominantly focused on statistical indicators of precipitation without discussing varying regional rainfall patterns, making it difficult to directly benefit flood safety management in meteorological and hydrological departments.

In light of the aforementioned challenges, this paper innovatively proposes the use of an unsupervised HDRC model for analysis of rainfall events in Kunming, Yunnan Province, China. This model aims to uncover the typical spatiotemporal distribution patterns of rainfall in this region in recent years and to conduct statistical indicator analysis.

2. Materials and Methods

This study establishes a standardized workflow for analyzing the spatiotemporal distribution characteristics of regional rainfall, including data cleaning, partitioning rainfall events, constructing high-dimensional learning samples, implementing an unsupervised dimensionality reduction-clustering models, and reconstructing the low-dimensional sample space. The specific process is illustrated in Figure 1.

2.1. Materials

This study utilizes rainfall data collected every 5 min from 62 rainfall stations in the urban area of Kunming City, provided by the Kunming Meteorological Bureau for the flood seasons from 2019 to 2021. These stations comprehensively cover the main built-up areas of Kunming City, with the distribution of the rain gauge stations illustrated in Figure 2.

Prior to data analysis, it is necessary to clean and screen historical rainfall data to eliminate unreasonable records:

- Records are considered unreasonable if a single station reports more than 10 mm of rainfall in 5 min without any rainfall 30 min before and after the event;
- Records are deemed unreasonable if a rainfall station within a 5 km × 5 km area reports 0 data, yet records more than 10 mm of rainfall in 5 min;
- For abnormal records at individual stations, rainfall isohyet maps for the period must be compared to verify the data's reasonableness. If found unreasonable, interpolation results from other stations within a 5 km × 5 km area of the station are used to replace its rainfall record.

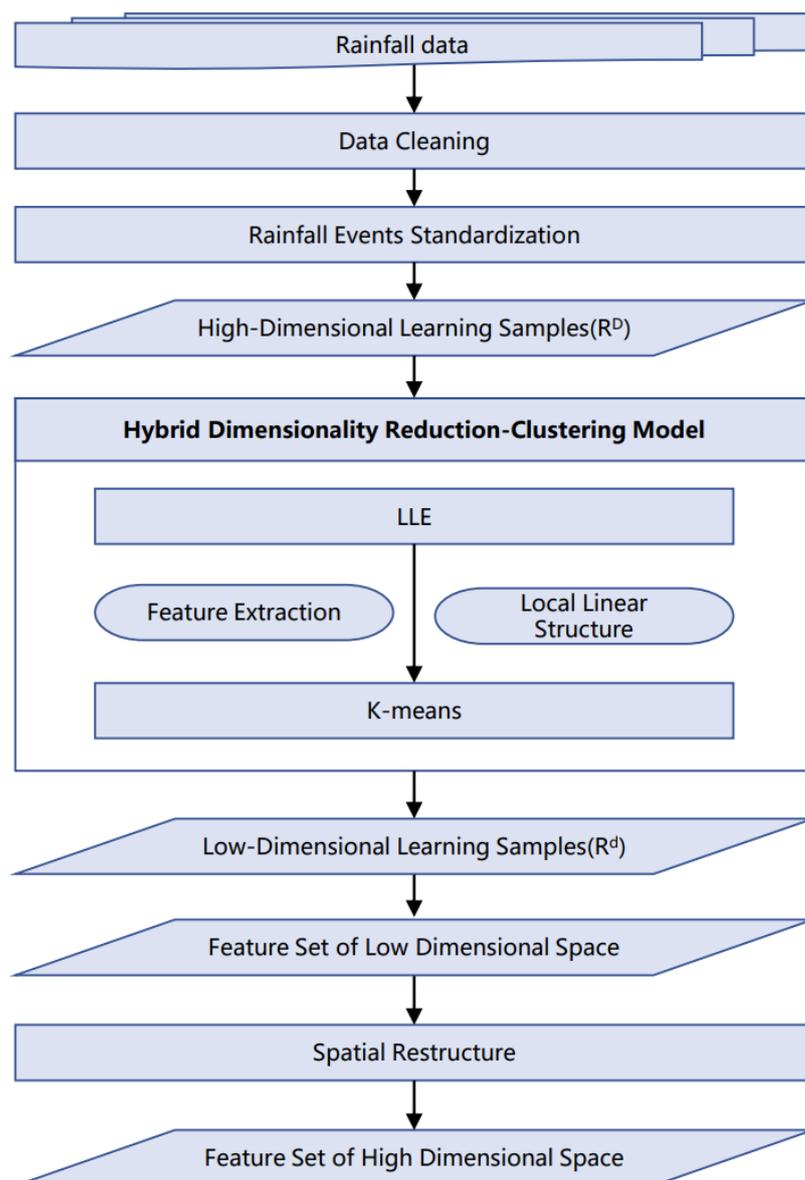


Figure 1. Flow chart of the analysis of spatial and temporal distribution characteristics of rainfall.

After initial cleaning of the rainfall data, specific rainfall events must be selected to reduce the sparsity of the dataset. To select as many standardized rainfall events as possible from the existing high-resolution data, the criteria for choosing rainfall events are as follows:

- Eliminate all periods in the annual time series of the study's rainfall stations where rainfall is consistently 0, considering discontinuities in the time series as the start of a new rainfall event;
- Temporally, eliminate rainfall events lasting less than one hour; volumetrically, exclude events where the average cumulative rainfall is less than 2 mm;
- Each rainfall event is downsampled to fit within a one-hour period (divided into twelve 5 min intervals), and the total volume is normalized to standardize the events.

Using the aforementioned methods, 161 rainfall events were selected between 2019 and 2021.

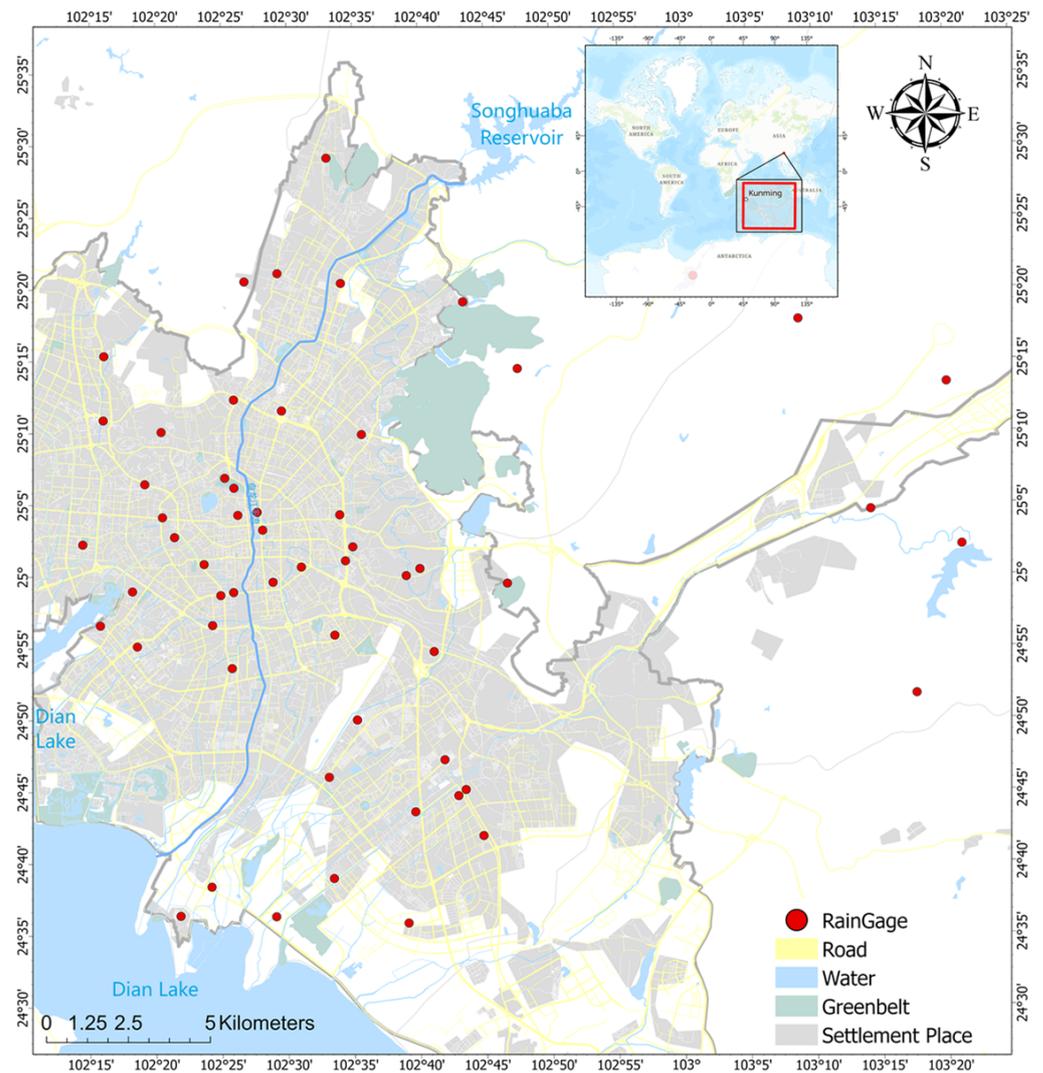


Figure 2. Study scope of the rainfall area.

2.2. Methods

2.2.1. Constructing High-Dimensional Data Samples

Following the data cleaning and normalization process, the dataset is transformed into a statistical learning sample set. This set contains the dynamic characteristics of the temporal and spatial dimensions of several rainfall events within the target area, as demonstrated in Equations (1) and (2):

$$\Omega = \{X_1, X_2, \dots, X_N\}, \tag{1}$$

$$X_i = \begin{bmatrix} H_{11} & H_{21} & \dots & H_{s1} \\ H_{12} & H_{22} & \dots & H_{s2} \\ \vdots & \vdots & & \vdots \\ H_{1m} & H_{2m} & \dots & H_{sm} \end{bmatrix}. \tag{2}$$

Here, Ω denotes the historical extreme rainfall sample set, comprising N extreme rainfall events. X_i represents the proportion matrix for the i th rainfall event at time t , and H_{it} denotes the percentage of rainfall at the i th rainfall station at time t out of the total rainfall at all stations at that time, as defined in Equation (3):

$$H_{it} = R_{it} / \sum_{i=1}^S R_{it}. \tag{3}$$

In the equation, R_{it} is the rainfall amount at station i at time t : $i = 1, 2, 3 \dots S$ (S is the number of rainfall stations), $t = 1, 2, 3 \dots m$ (m is the number of time intervals).

Ω represents a high-dimensional data space describing the rainfall processes in a specific region. To enhance computational efficiency and reduce redundancy in high-dimensional data—thus preventing confusion of the main components in the rainfall sample space via clustering algorithms—it is assumed that the rainfall sample vectors of an area with specific climatic features, hydrological distribution, and topography are distributed within a specific low-dimensional manifold. Subsequent dimensionality reduction is performed using the manifold learning algorithm to be introduced.

2.2.2. Unsupervised HDRC Models

The unsupervised HDRC model developed in this study comprises two components: first, the application of the Locally Linear Embedding (LLE) algorithm from manifold learning for dimensionality reduction of high-dimensional data samples; second, the implementation of the dynamic clustering K-Means algorithm for clustering analysis of the low-dimensional data samples.

- LLE

The LLE algorithm is a widely utilized unsupervised technique for dimensionality reduction, designed to map high-dimensional data into a lower-dimensional space without altering its intrinsic structure [39]. The primary goal of LLE is to preserve the local structure of the original data as much as possible, discarding irrelevant noise and the global structure through dimensionality reduction. The fundamental concept of LLE involves finding the K nearest neighbors for each data point and representing that point as a linear combination of these neighbors. This process calculates the low-dimensional representation by minimizing the distance between these linear combinations and the original data points. LLE stands out as an effective dimensionality reduction tool that maintains the local structural integrity of the data, reduces redundancy present in high-dimensional spaces, and operates efficiently without the need for any prior knowledge or intricate hyperparameter adjustments [40]. The LLE algorithm is implemented in three main steps:

1. In high-dimensional space, the LLE algorithm identifies the K nearest neighbors of a sample x_i using the Euclidean distance metric.

Firstly, for N data points $\{x_1, x_2, \dots, x_N\} \in R^D$ in the high-dimensional space, calculate the Euclidean distance between each sample point x_i and all other samples as demonstrated in Equation (4):

$$x_i \approx \bar{x}_i = \sum_{j=1}^k w_{ij}x_j, \tag{4}$$

subject to the constraints in Equation (5):

$$\sum_{j=1}^k w_{ij} = 1. \tag{5}$$

2. For each sample x_i , find the linear relationship of the K nearest neighbors in its neighborhood to obtain the linear relationship weight coefficients, as shown in Equation (6):

$$W_i = (w_{i1}, w_{i2}, \dots, w_{ik})^T (i = 1, 2, \dots, N). \tag{6}$$

Next, by solving Equation (4) at the minimum of the constraint of Equation (5), the weight coefficients W are obtained, as shown in Equation (7):

$$W_i = \frac{Z_i^{-1}I}{I^T Z_i^{-1} I}, \tag{7}$$

The weight vector $W_i = (\mathbf{w}_{i1}, \mathbf{w}_{i2}, \dots, \mathbf{w}_{ik})^T (i = 1, 2, \dots, N)$ is obtained; in Equation (7), Z_i^{-1} is $k \times k$ dimensional, I is $k \times 1$ dimensional, and the coefficient matrix W is $N \times k$ dimensional.

- Assuming the linear relationship weight coefficients W_i within the K neighborhood remain constant between high-dimensional and low-dimensional spaces, the sample data are reconstructed in the lower dimension using the weight coefficients W_i , and the following are implemented: $x_i \in R^D \rightarrow y_i \in R^d, d \ll D$.

The LLE algorithm assumes that mapping high-dimensional samples to a lower-dimensional space preserves the local linear relationships of samples in the high-dimensional space with constant weight coefficients. Then, the point $\{x_1, x_2, \dots, x_N\} \in R^D$ in the high-dimensional space is mapped into the low-dimensional space as $\{y_1, y_2, \dots, y_N\} \in R^d$ through the weight coefficients W . The mapping x_i of the high-dimensional sample point y_i in the low-dimensional space can be similarly obtained by solving for the minimum of the mean square deviation; the objective function is given by Equation (8):

$$\operatorname{argmin}_y \sum_{i=1}^N \left\| y_i - \sum_{j=1}^k w_{ij} y_j \right\|_2^2. \tag{8}$$

In order to obtain normalized low-dimensional data, the constraints are as shown in Equation (9):

$$\text{s.t. } \sum_{i=1}^N y_i = 0; \frac{1}{N} \sum_{i=1}^N y_i y_i^T = I \tag{9}$$

where $y_i \in R^d (d \ll D)$ is solved by making the points outside the neighborhood of i $w_{ij} = 0$, thus expanding the matrix of weight coefficients from $W_{N \times K}$ to the square matrix $W_{N \times N}$.

The mean square deviation is written in the following form, as shown in Equation (10):

$$G(Y) = \sum_{i=1}^N \left\| y_i - \sum_{j=1}^N w_{ij} y_j \right\|_2^2 = \operatorname{tr} \left(Y(I - W)(I - W)^T Y^T \right). \tag{10}$$

Let $M = (I - W)(I - W)^T$ and Equation (10) becomes Equation (11):

$$G(Y) = \operatorname{tr} \left(YMY^T \right). \tag{11}$$

Using the same method, Equation (11) is transformed into Equation (12) by synthesizing an optimization function using the Lagrange operator

$$G(Y) = \sum_{i=1}^N \left\| y_i - \sum_{j=1}^N w_{ij} y_j \right\|_2^2 = \operatorname{tr} \left(Y(I - W)(I - W)^T Y^T \right). \tag{12}$$

Derive $H(Y)$ and make it 0 as in Equation (13):

$$\frac{\partial H(Y)}{\partial Y} = MY^T + \lambda I = 0. \tag{13}$$

It follows that Y is a matrix consisting of the eigenvectors of M . In order to reduce the data to d dimensions, the eigenvectors corresponding to the first d non-zero eigenvalues of the matrix M are taken to be the desired low-dimensional mapping Y .

- K-Means**

K-Means is a clustering analysis algorithm that divides a set of unlabeled data points into distinct groups or “clusters”, defined by the similarity or proximity of data points within each cluster [41]. The fundamental concept of this algorithm is to assign data points to a predetermined number of clusters to minimize the distance between each data point and the centroid of its cluster while maximizing the distance between different clusters [42].

The process involves partitioning n data points into k clusters, where k is a predetermined constant. Clusters are defined based on the similarity or proximity of the data points within them, measured using metrics such as Euclidean distance, Manhattan distance, or cosine similarity. In K-Means, each cluster is characterized by a centroid, which is the mean of all data points within the cluster. The algorithm begins by randomly selecting k data points as the initial centroids and then iterates by assigning the remaining data points to the nearest cluster based on distance, updating the centroids of the clusters and reassigning data points until the allocation stabilizes or a set number of iterations is reached. Finally, the K-Means algorithm will return k clusters and their centroid of mass [43]. The selection of the hyperparameter n is critical and can be initially estimated based on prior knowledge, with adjustments made by evaluating the similarity of clustering outcomes to avoid misclassification [44].

The above method is able to find the subset of rainfall feature space clustering after applying the stream learning dimensionality reduction algorithm, as shown in Equation (14):

$$C = \{C_1, C_2, \dots, C_r\}. \quad (14)$$

These clusters, while lacking physical dimensionality, represent the projection of the original sample feature space onto a low-dimensional manifold. The LLE algorithm presupposes that local linear relationships within high-dimensional space are preserved in the low-dimensional space, meaning the linear relationship between a sample x_i and its neighbors in high-dimensional space remains consistent with its projection y_i and corresponding neighbors in low-dimensional space. Thus, the clusters $C = \{C_1, C_2, \dots, C_r\}$ obtained through dynamic clustering are applicable in the physically meaningful high-dimensional sample space. By reverse-mapping these clusters using sample indices, clusters $B = \{B_1, B_2, \dots, B_r\}$ in high-dimensional space are identified, enabling the analysis of spatial and temporal distribution characteristics of rainfall in high-dimensional space, as shown in Equation (15):

$$S_j = \frac{1}{B_j} \sum_{x_i \in B_j} x_i \in \mathbb{R}^D \quad (15)$$

3. Results

Given the rainfall data for Kunming City and the computational process introduced in section two, utilizing the elbow method, combined with the prior knowledge of local meteorological and hydrological departments, three types of rainfall events were identified from a total of 161 rainfall occurrences. The clustering visualization is presented in Figure 3, where Feature Dimension 1 and Feature Dimension 2 represent low-dimensional components obtained through the LLE algorithm and lack specific physical meaning. These dimensions are abstract representations, capturing the essence of the data's structure in a reduced form to facilitate the clustering and visualization of the underlying patterns within the rainfall events.

The statistical information for the three types of rainfall is shown in Table 1. To spatially resample these three characteristic rainfall types, the Kriging interpolation method was employed. The spatial and temporal distribution maps of the rainfall are depicted in Figures 4–6.

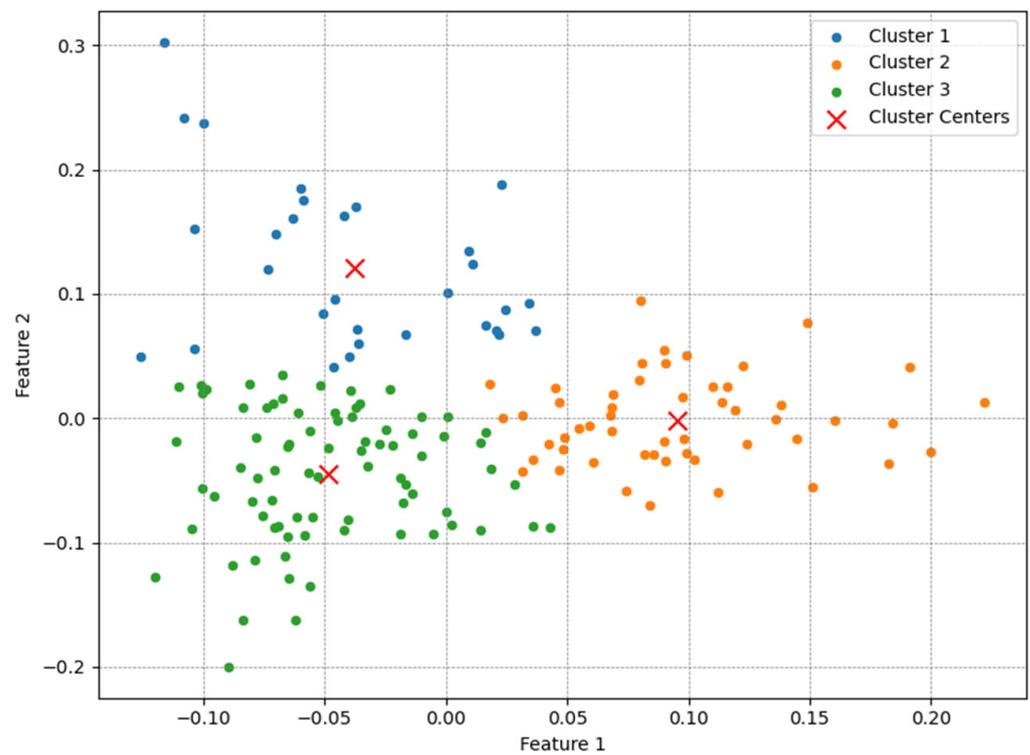


Figure 3. Low-dimensional sample set clustering visualization.

Table 1. Statistical characteristics of various rainfall types.

Type	Month	Number of Rainfall Events	Duration	Average Surface Rainfall [mm]	Maximum Hourly Rainfall Intensity at the Center of the Rainfall [mm/h]
1	5	8	3.45	6.21	18.45
	6	15	4.95	14.46	30.00
	7	21	6.78	16.81	31.50
	8	19	4.28	9.28	26.67
	9	16	4.61	11.13	23.54
2	5	1	1.33	4.57	11.00
	6	6	4.59	13.17	29.67
	7	7	3.58	6.17	16.24
	8	12	5.12	15.31	33.14
	9	4	3.65	5.50	17.28
3	5	3	4.14	10.15	19.73
	6	14	4.47	14.55	28.89
	7	14	4.79	11.06	19.89
	8	11	3.78	7.60	23.53
	9	10	2.92	6.01	19.53

Type I rainfall is characterized by rain centers distributed along the northern shore of Dian Lake and the southern part of Kunming’s main urban area. As observed in Figure 4, the spatial dynamics of this rainfall type exhibit a progression from the waters of Dian Lake gradually towards the land, posing a primary threat to the western and southern regions of Kunming’s main urban area. According to Table 1, this type of rainfall accounts for 79 events, making up 49.07% of the study samples and representing the predominant rainfall pattern in the Kunming area. These rainfall events mainly occur in June, July, August, and September, with an average duration of 4.81 h. The maximum hourly rainfall intensity at the center reaches 26.0 mm/h. The average rainfall duration peaks in July at 6.78 h, and the average rainfall amount is 16.81 mm.

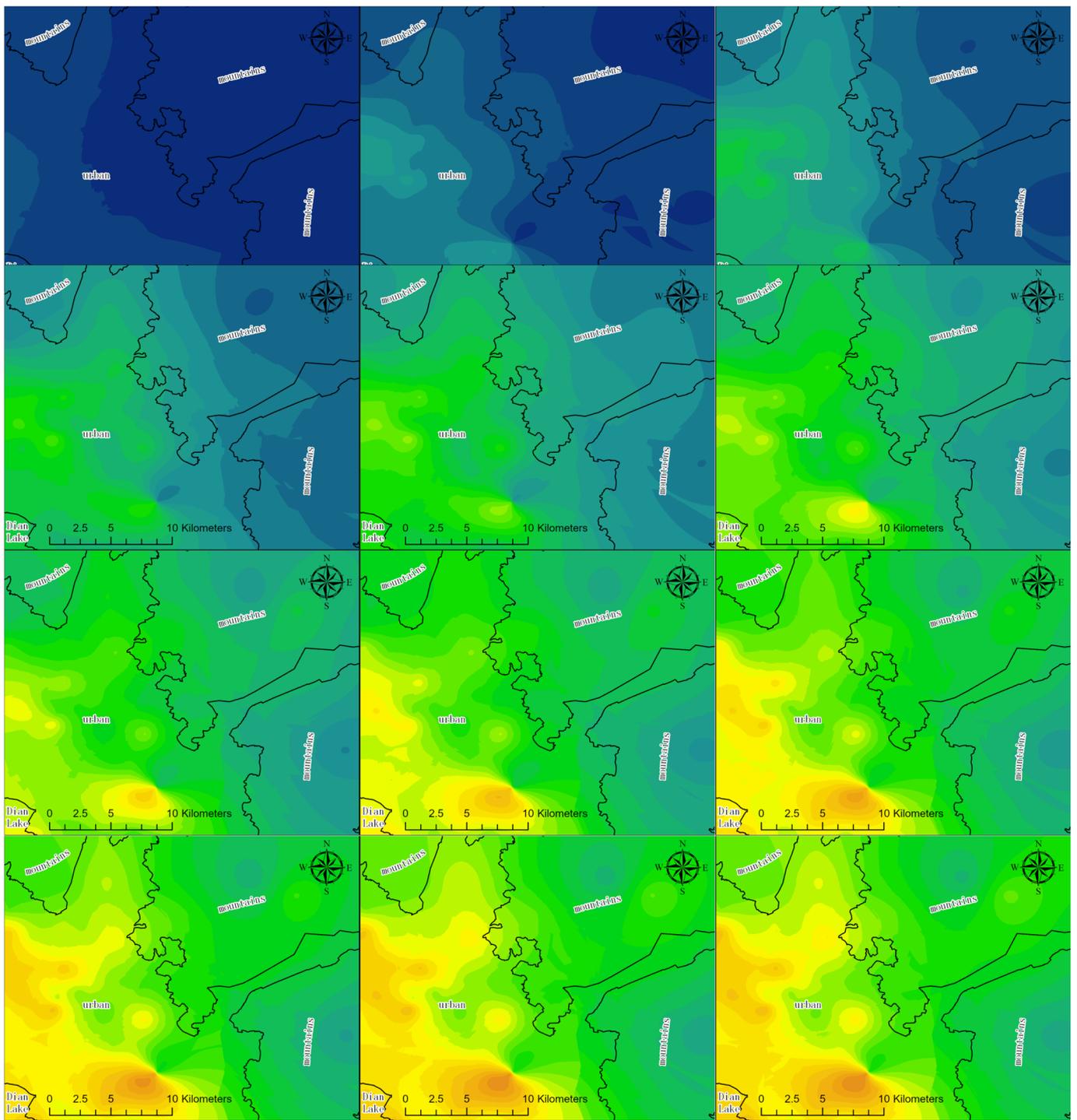


Figure 4. Cumulative rainfall distribution for Type I rainfall.

Type II rainfall is characterized by its concentration in the northern mountainous region of Kunming City. As depicted in Figure 5, the spatial dynamic change of this rainfall type is relatively minor, with a more fixed and concentrated rainfall center due to the orographic effect of the mountain’s windward slopes. The distribution of the rainfall center aligns closely with the contour lines of the mountain, posing a significant threat to the northern mountainous areas and urban regions of Kunming, potentially triggering flash floods and affecting the northern urban areas. According to Table 1, this type of rainfall primarily occurs during the hottest months, July and August, with a higher frequency

in August. Among the three rainfall types, it accounts for the smallest proportion of the study sample with 30 events. However, it features the highest maximum hourly rainfall intensity at 33.14 mm/h, marking it as a type of rainfall with highly uneven spatial and temporal distribution.

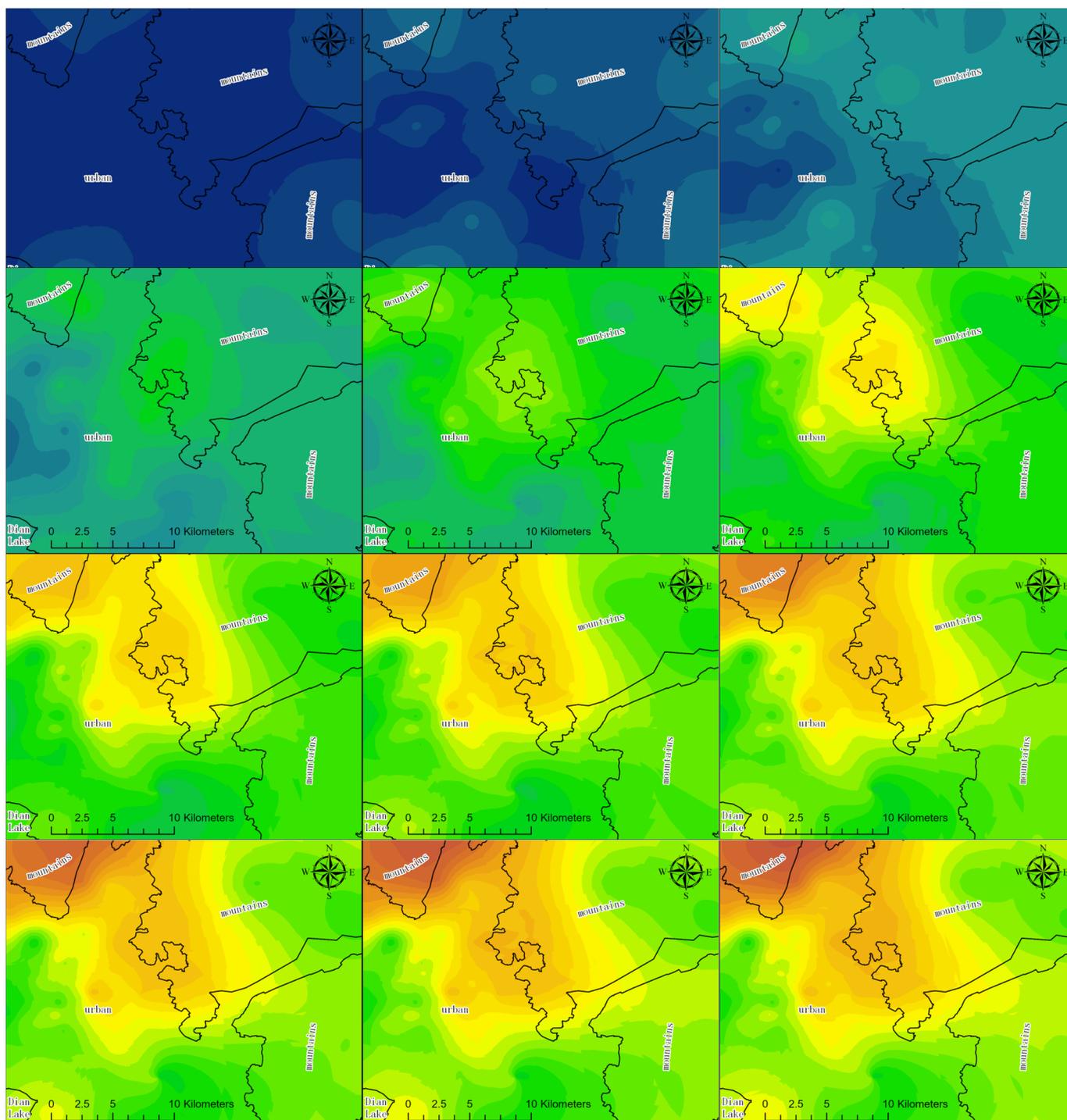


Figure 5. Cumulative rainfall distribution for Type II rainfall.

Type III rainfall is centered around the main urban area of Kunming City and is associated with the city's urban heat island effect. According to Table 1, this type of rainfall shows relatively minor variations in its parameters and primarily occurs during the cooler months of May and September, with maximum hourly rainfall intensities of 19.73 mm/h

and 19.53 mm/h, respectively. This type of rainfall mainly takes place within the urban boundaries and is a significant cause of urban flooding in Kunming. Figure 6 reveals that there is a rainfall accumulation in the lowland at the intersection of the central four districts of Kunming City, indicating a specific area where rainfall tends to accumulate, potentially exacerbating flooding issues within the urban environment.

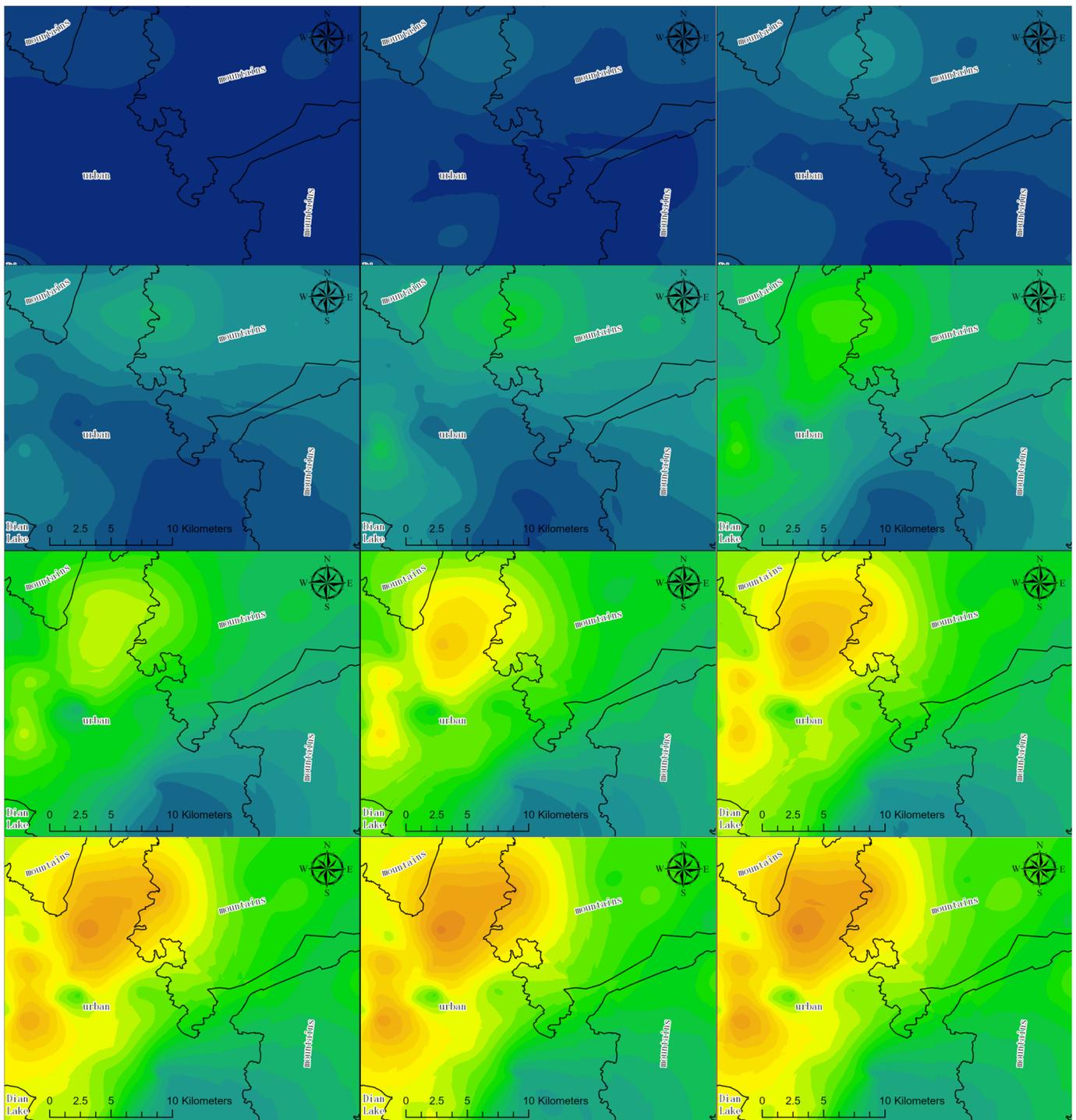


Figure 6. Cumulative rainfall distribution for Type III rainfall.

4. Discussion

To further cross-validate the correlation of the temporal distribution of rainfall among the three different types of rainfall across various regions, this paper categorizes the rain gauge stations within the study area into three types based on their locations: the first type consists of rain gauge stations located in urban areas, the second type includes stations close to water bodies (Dian Lake), and the third type comprises stations in mountainous areas. We statistically analyzed the maximum cumulative proportion and the average cumulative proportion of different rainfall types at the rain gauge stations in different regions, as shown in Figure 7. Figure 7a corresponds to Type I rainfall, where the slope of the curves for all three regions is similar, indicating a relatively even distribution of rainfall, consistent with the prior knowledge that water vapor from the water body enters the plain area. Figure 7b pertains to Type II rainfall, which demonstrates that the rainfall in the mountainous area is the greatest, while that in the urban area is the least. Although the mountainous area has an advantage in the maximum rainfall amount, its average rainfall is consistent with that near Dian Lake, suggesting that the rainfall in the mountainous area is short-lived but intense, aligning with the results presented in section three. Type III rainfall corresponds to urban area rainfall; as seen in Figure 7c, this type of rainfall predominates in urban areas, illustrating the stimulatory effect of the urban heat island effect on rainfall events.

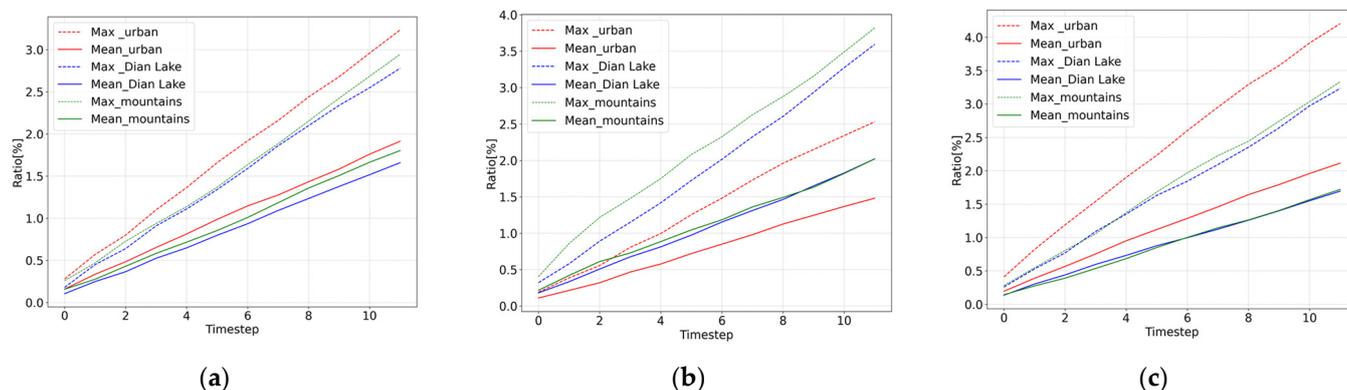


Figure 7. Cumulative rainfall ratios for different regions for three different types of rainfall: (a) Type I rainfall; (b) Type II rainfall; and (c) Type III rainfall.

5. Conclusions

This chapter analyzes rainfall data from 61 stations in Kunming City from 2019 to 2021, cleaning and vetting the data at 5 min intervals. After screening, 161 rainfall events were identified, and a rainfall sample database for Kunming City was constructed. Utilizing manifold learning and dynamic clustering algorithms, an unsupervised HDRC model was developed to analyze the spatiotemporal distribution characteristics of rainfall in Kunming City. The rainfall was categorized into three types based on these characteristics. Analyses and summaries were conducted on the spatiotemporal distribution, occurrence, timing, frequency, and potential disaster-prone area characteristics of each type of rainfall. This paper fills in gaps in the use of unsupervised learning methods for classifying meteorological data and identifying extreme rainfall events. Exploration of rainfall patterns in the target region was conducted from multiple perspectives, including the temporal distribution, spatial patterns, and statistical analysis of classification results. The overview of the three types of rainfall is as follows:

1. Type I rainfall is characterized by a distribution center located between the northern shore of Dian Lake and the southern part of Kunming's main urban area. Rainfall distribution is relatively uniform, posing a primary threat to the western and southern regions of Kunming's main urban area. This type of rainfall comprises 79 events, accounting for 49.07% of the study sample, and is the predominant rainfall type in the Kunming area. It primarily occurs in June, July, August, and September.

2. Type II rainfall has its center in the northern mountainous region of Kunming, with minimal spatial dynamic variation. Among the rainfall types, this class, comprising 30 events and accounting for 18.63% of the sample, is the least common. It predominantly occurs in July and August, the warmest months, with August witnessing the highest frequency and peak hourly intensity of 33.14 mm/h, indicating a highly uneven spatiotemporal distribution of precipitation.

3. The precipitation center of Type III rainfall is located in the urban area of Kunming, characterized by and associated with the urban rain island effect. Variability in rainfall parameters is minimal. This category consists of 52 events, representing 32.30% of the study sample, predominantly occurring in the cooler months of May and September, with peak hourly intensities of 19.73 mm/h and 19.53 mm/h, respectively. This type of rainfall predominantly occurs within urban areas and is the primary cause of urban waterlogging in Kunming.

The unsupervised HDRC model enables rapid classification of rainfall processes in a region, producing various types of typical rainfall events and their characteristic values, thereby facilitating a deeper understanding of the influence of topography, hydrology, and other factors on the local hydrological cycle. In practical applications, predictions of the most likely rainfall types can be made based on the timing of rainfall events and factors such as the direction of moisture sources provided by meteorological forecasts. This allows for the implementation of targeted flood and flood control measures such as reinforcing flood prevention arrangements at rainfall centers. This approach has potential for application in cities with varied climate characteristics, water–land distributions, and topographical features, offering valuable decision-making references for local meteorological and flood control departments.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L. and H.R.; Software, N.Z. and M.W.; Validation, Y.W.; Formal analysis, W.L. and H.R.; Investigation, J.W.; Resources, M.L.; Data curation, M.L.; Writing—original draft, W.L.; Writing—review & editing, N.L. and M.W.; Supervision, N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Yangtze River Water Science Joint Fund (U2240203) and National Key Research and Development Program of China (2022YFC3006400, 2022YFC3006403, 2023YFC3008502).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data involved in the study are primarily confidential to the study area and require departmental approval before they can be accessed.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meehl, G.A.; Zwiers, F.; Evans, J.; Knutson, T.; Mearns, L.; Whetton, P. Trends in Extreme Weather and Climate Events: Issues Related to Modeling Extremes in Projections of Future Climate Change. *Bull. Am. Meteorol. Soc.* **2000**, *81*, 427–436. [[CrossRef](#)]
2. Tabari, H. Climate Change Impact on Flood and Extreme Precipitation Increases with Water Availability. *Sci. Rep.* **2020**, *10*, 13768. [[CrossRef](#)] [[PubMed](#)]
3. Huong, H.T.L.; Pathirana, A. Urbanization and Climate Change Impacts on Future Urban Flooding in Can Tho City, Vietnam. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 379–394. [[CrossRef](#)]
4. Alcamo, J.; Flörke, M.; Märker, M. Future Long-Term Changes in Global Water Resources Driven by Socio-Economic and Climatic Changes. *Hydrol. Sci. J.* **2007**, *52*, 247–275. [[CrossRef](#)]
5. Lobell, D.B.; Gourdj, S.M. The Influence of Climate Change on Global Crop Productivity. *Plant Physiol.* **2012**, *160*, 1686–1697. [[CrossRef](#)] [[PubMed](#)]
6. Leal Filho, W.; Totin, E.; Franke, J.A.; Andrew, S.M.; Abubakar, I.R.; Azadi, H.; Nunn, P.D.; Ouweneel, B.; Williams, P.A.; Simpson, N.P. Understanding Responses to Climate-Related Water Scarcity in Africa. *Sci. Total Environ.* **2022**, *806*, 150420. [[CrossRef](#)] [[PubMed](#)]

7. Brunner, M.I.; Slater, L.; Tallaksen, L.M.; Clark, M. Challenges in Modeling and Predicting Floods and Droughts: A Review. *Wiley Interdiscip. Rev. Water* **2021**, *8*, e1520. [[CrossRef](#)]
8. Trenberth, K.E.; Dai, A.; Rasmussen, R.M.; Parsons, D.B. The Changing Character of Precipitation. *Bull. Am. Meteorol. Soc.* **2003**, *84*, 1205–1218. [[CrossRef](#)]
9. Bales, R.C.; Molotch, N.P.; Painter, T.H.; Dettinger, M.D.; Rice, R.; Dozier, J. Mountain Hydrology of the Western United States. *Water Resour. Res.* **2006**, *42*, W08432. [[CrossRef](#)]
10. Paschalis, A.; Fatichi, S.; Molnar, P.; Rimkus, S.; Burlando, P. On the Effects of Small Scale Space–Time Variability of Rainfall on Basin Flood Response. *J. Hydrol.* **2014**, *514*, 313–327. [[CrossRef](#)]
11. Wheeler, H.S.; Chandler, R.E.; Onof, C.J.; Isham, V.S.; Bellone, E.; Yang, C.; Lekkas, D.; Lourmas, G.; Segond, M.-L. Spatial-Temporal Rainfall Modelling for Flood Risk Estimation. *Stoch. Environ. Res. Risk Assess.* **2005**, *19*, 403–416. [[CrossRef](#)]
12. Fatichi, S.; Vivoni, E.R.; Ogden, F.L.; Ivanov, V.Y.; Mirus, B.; Gochis, D.; Downer, C.W.; Camporese, M.; Davison, J.H.; Ebel, B. An Overview of Current Applications, Challenges, and Future Trends in Distributed Process-Based Models in Hydrology. *J. Hydrol.* **2016**, *537*, 45–60. [[CrossRef](#)]
13. Panagoulia, D.; Bárdossy, A.; Lourmas, G. Diagnostic Statistics of Daily Rainfall Variability in an Evolving Climate. *Adv. Geosci.* **2006**, *7*, 349–354. [[CrossRef](#)]
14. Trenberth, K.E. The Impact of Climate Change and Variability on Heavy Precipitation, Floods, and Droughts. *Encycl. Hydrol. Sci.* **2005**, *17*, 1–11.
15. Ramírez, J.A. Prediction and Modeling of Flood Hydrology and Hydraulics. In *Inland Flood Hazards: Human, Riparian and Aquatic Communities*; Cambridge University Press: Cambridge, UK, 2000; p. 498.
16. Tapiador, F.J.; Turk, F.J.; Petersen, W.; Hou, A.Y.; García-Ortega, E.; Machado, L.A.; Angelis, C.F.; Salio, P.; Kidd, C.; Huffman, G.J. Global Precipitation Measurement: Methods, Datasets and Applications. *Atmos. Res.* **2012**, *104*, 70–97. [[CrossRef](#)]
17. Rahman, M.A.; Yunsheng, L.; Sultana, N. Analysis and Prediction of Rainfall Trends over Bangladesh Using Mann–Kendall, Spearman’s Rho Tests and ARIMA Model. *Meteorol. Atmos. Phys.* **2017**, *129*, 409–424. [[CrossRef](#)]
18. Borges, P.d.A.; Franke, J.; da Anunciação, Y.M.T.; Weiss, H.; Bernhofer, C. Comparison of Spatial Interpolation Methods for the Estimation of Precipitation Distribution in Distrito Federal, Brazil. *Theor. Appl. Climatol.* **2016**, *123*, 335–348. [[CrossRef](#)]
19. Lima, A.O.; Lyra, G.B.; Abreu, M.C.; Oliveira-Júnior, J.F.; Zeri, M.; Cunha-Zeri, G. Extreme Rainfall Events over Rio de Janeiro State, Brazil: Characterization Using Probability Distribution Functions and Clustering Analysis. *Atmos. Res.* **2021**, *247*, 105221. [[CrossRef](#)]
20. Marquínez, J.; Lastra, J.; García, P. Estimation Models for Precipitation in Mountainous Regions: The Use of GIS and Multivariate Analysis. *J. Hydrol.* **2003**, *270*, 1–11. [[CrossRef](#)]
21. Atluri, G.; Karpatne, A.; Kumar, V. Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Comput. Surv.* **2018**, *51*, 1–41. [[CrossRef](#)]
22. Cressie, N.; Wikle, C.K. *Statistics for Spatio-Temporal Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015; ISBN 1-119-24304-1.
23. Karpatne, A.; Ebert-Uphoff, I.; Ravela, S.; Babaie, H.A.; Kumar, V. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1544–1554. [[CrossRef](#)]
24. Pulukool, F.; Li, L.; Liu, C. Using Deep Learning and Machine Learning Methods to Diagnose Hailstorms in Large-Scale Thermodynamic Environments. *Sustainability* **2020**, *12*, 10499. [[CrossRef](#)]
25. Schultz, M.G.; Betancourt, C.; Gong, B.; Kleinert, F.; Langguth, M.; Leufen, L.H.; Mozaffari, A.; Stadler, S. Can Deep Learning Beat Numerical Weather Prediction? *Philos. Trans. R. Soc. A* **2021**, *379*, 20200097. [[CrossRef](#)]
26. Salcedo-Sanz, S.; Ghamisi, P.; Piles, M.; Werner, M.; Cuadra, L.; Moreno-Martínez, A.; Izquierdo-Verdiguier, E.; Muñoz-Marí, J.; Mosavi, A.; Camps-Valls, G. Machine Learning Information Fusion in Earth Observation: A Comprehensive Review of Methods, Applications and Data Sources. *Inf. Fusion* **2020**, *63*, 256–272. [[CrossRef](#)]
27. Dhall, D.; Kaur, R.; Juneja, M. Machine Learning: A Review of the Algorithms and Its Applications. In *Proceedings of ICRIC 2019: Recent Innovations in Computing*; Springer: Cham, Switzerland, 2020; pp. 47–63.
28. Hiran, K.K.; Jain, R.K.; Lakhwani, K.; Doshi, R. *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*; BPB Publications: Noida, India, 2021; ISBN 93-91392-35-0.
29. Nanga, S.; Bawah, A.T.; Acquaye, B.A.; Billa, M.-I.; Baeta, F.D.; Odai, N.A.; Obeng, S.K.; Nsiah, A.D. Review of Dimension Reduction Methods. *J. Data Anal. Inf. Process.* **2021**, *9*, 189–231. [[CrossRef](#)]
30. Anowar, F.; Sadaoui, S.; Selim, B. Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (Pca, Kpca, Lda, Mds, Svd, Lle, Isomap, Le, Ica, t-Sne). *Comput. Sci. Rev.* **2021**, *40*, 100378. [[CrossRef](#)]
31. Sisodia, D.; Singh, L.; Sisodia, S.; Saxena, K. Clustering Techniques: A Brief Survey of Different Clustering Algorithms. *Int. J. Latest Trends Eng. Technol. (IJLTET)* **2012**, *1*, 82–87.
32. Celebi, M.E.; Aydin, K. *Unsupervised Learning Algorithms*; Springer: Cham, Switzerland, 2016; Volume 9.
33. Xu, R.; Wunsch, D. Survey of Clustering Algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [[CrossRef](#)]
34. Liu, Y.-Y.; Li, L.; Liu, Y.-S.; Chan, P.W.; Zhang, W.-H. Dynamic Spatial-Temporal Precipitation Distribution Models for Short-Duration Rainstorms in Shenzhen, China Based on Machine Learning. *Atmos. Res.* **2020**, *237*, 104861. [[CrossRef](#)]
35. Oppel, H.; Fischer, S. A New Unsupervised Learning Method to Assess Clusters of Temporal Distribution of Rainfall and Their Coherence with Flood Types. *Water Resour. Res.* **2020**, *56*, e2019WR026511. [[CrossRef](#)]

36. Mercer, A.; Dyer, J. A Southeastern United States Warm Season Precipitation Climatology Using Unsupervised Learning. *Climate* **2022**, *11*, 2. [[CrossRef](#)]
37. Wang, L.; Shi, Z.; Ye, L.; Su, B. Analysis on the Characteristics of Extreme Weather Events in Kunming City during Recent 20 Years. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *252*, 042124. [[CrossRef](#)]
38. Zhang, Y.; Luo, Y.; Su, H. Study on Land Use and Precipitation Changes in the Main Urban Area of Kunming in the Past 40 Years. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *658*, 012008. [[CrossRef](#)]
39. Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326. [[CrossRef](#)]
40. Chang, H.; Yeung, D.-Y. Robust Locally Linear Embedding. *Pattern Recognit.* **2006**, *39*, 1053–1065. [[CrossRef](#)]
41. Shi, N.; Liu, X.; Guan, Y. Research on K-Means Clustering Algorithm: An Improved k-Means Clustering Algorithm. In Proceedings of the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, 2–4 April 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 63–67.
42. Žalik, K.R. An Efficient K'-Means Clustering Algorithm. *Pattern Recognit. Lett.* **2008**, *29*, 1385–1391. [[CrossRef](#)]
43. Hamerly, G.; Elkan, C. Learning the k in K-Means. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2003; Volume 16.
44. Krishna, K.; Murty, M.N. Genetic K-Means Algorithm. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **1999**, *29*, 433–439. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.