J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 5453--5470

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on pattern recognition and spatial prediction method of lightning activity distribution based on cluster analysis

Amuersana[1], Shi Jin[1,✉], Lu Chao[1], Xuan Li[1], Danping Wang[1]

[1] *Meteorological Disaster Prevention Center, Hohhot Meteorological Bureau, Hohhot, Inner Mongolia, 010010, China*

ABSTRACT

Frequent lightning activity has the potential to cause damage to man-made facilities, cause forest fires and other hazards, and the prediction of lightning activity can help to avoid the occurrence of these disasters. In this paper, based on the lightning activity data of a region, the distribution pattern of lightning activity is identified at different elevations and latitudes and longitudes. Then geodetic distance and contributing nearest-neighbor similarity are introduced, and a GS-DBSCAN clustering algorithm is proposed to realize the spatial prediction of lightning activity by using the method of least-squares fitting of prediction equations. The lightning activity directions after data clustering show topographic correlation, and the overlap between lightning activity directions and topography is about 35%. Combined with the prediction images, it is found that the lightning activity prediction results of this paper's method are closer to the real value than other algorithms, with an average offset error of less than 1.1km, an accuracy rate of >85%, and a false alarm rate of <35%, which reflects a good prediction performance.

Keywords: GS-DBSCAN, cluster analysis, least squares, lightning activity, spatial prediction

1. Introduction

Lightning is a very spectacular ultra-long distance discharge process in the atmosphere, usually accompanied by strong convective weather processes, is a common natural phenomenon [1]. Lightning is capable of producing great damage in an instant due to its physical effects such as powerful electrical currents, scorching heat, intense electromagnetic radiation, and violent shock waves, resulting in lightning disasters. Lightning strikes often lead to casualties, and may also cause damage to buildings, power supply and distribution systems, communication equipment, and civil electrical appliances, forest fires, and combustion or even explosions in warehouses, refineries, and oil

✉ Corresponding author.
E-mail address: messidiegoks@126.com (S. Jin).
Received 15 January 2024; Revised 10 March 2024; Accepted 30 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-305

fields, resulting in significant economic losses and adverse social impacts [2-3]. Therefore, lightning disaster is a serious threat that cannot be ignored for the safety production of all sectors in China. Lightning disaster prevention and mitigation work appears to be extremely urgent and important, and lightning disaster prevention and mitigation work must be established on the basis of mastering the local lightning activity law and reasonable lightning disaster prevention and mitigation planning [4-5]. In order to effectively carry out lightning disaster defense work, especially in order to solve the problems of urban planning, project siting and functional zoning layout, it is crucial to identify and spatially predict the lightning distribution pattern starting from the historical data of monitoring and observation [6-7]. Among them, the problem of how to more accurately describe and predict the direction and coverage area of thunderstorm clouds can provide a new solution to the study of the location of lightning strikes, probability and calculation of the size of the lightning current.

Compared with the traditional lightning arrester and lightning line methods, some scholars have considered the spatial and temporal distribution of lightning and capital investment. Literature [8] investigated the relationship between environmental variables and lightning occurrence, and utilized sensors to collect environmental variables to construct a global climate model, which was used to predict the occurrence of lightning, showing good potential for application. Literature [9] developed a localized lightning forecasting system, which uses a machine learning model to learn the weather values in recent years as well as the lightning activity records, and calculates and formulates the lightning occurrence thresholds, which can achieve lightning forecasting in different environments. Literature [10] constructs a weather research and forecasting model with electrification add-on package (WRF-ELEC) and compares the model's lightning activity prediction performance under different data sources. Literature [11] describes the application path of deep learning techniques in lightning prediction, combining computer vision and spatio-temporal data analysis on the basis of numerical prediction to build powerful lightning prediction tools to cope with the increasingly frequent extreme weather changes. Literature [12] establishes a weather forecast parameter model based on machine learning technology, which can provide early warning of upcoming lightning activities by collecting meteorological data from observation points, and further integrates data mining technology to extract the characteristic parameters in complex data, which can realize the prediction of spatio-temporal distribution patterns of lightning activities. Literature [13] used artificial neural networks to establish lightning prediction models based on unbalanced data and balanced data respectively, the latter was balanced by synthetic oversampling techniques, and both models showed good lightning prediction accuracy in case studies. Literature [14] designed a random forest algorithm for proximity prediction and very short-term prediction, which was applied to process high-resolution thermodynamic parameters generated by numerical weather prediction, etc., and was able to improve the accuracy of lightning prediction in the region. Literature [15] uses spectral satellite images and lightning activity records as data sources, and employs an end-to-end deep convolutional neural network to construct a lightning prediction model, and obtains accurate lightning prediction results by dealing with the imbalance between positive and negative data through appropriate methods. Although the above methods can track and predict more accurately, the methods mostly study the thundercloud changes and do not pay enough attention to the data and extinction of the falling lightning in reality. After dividing the number of falling mines into the density of falling mines by grid, the trend change and spatial prediction performance of thunderstorm clouds are greatly enhanced by clustering analysis of the relationship between the density of falling mines between each grid and its neighbors.

The article first identifies the distribution characteristics of lightning activity under different elevation, longitude and latitude conditions by taking the statistics of lightning activity in a region of China as an example, and then proposes an improved clustering algorithm and a spatial prediction method of lightning activity by the least squares method. The classical DBSCAN density clustering method is optimized by using geodetic distance instead of Euclidean distance and introducing the shared nearest neighbor similarity SNN to represent the clustering density. The Eps and MinPts parameters are adaptively determined according to the distribution characteristics of the data, and the prediction of lightning activity is realized by choosing the method of least squares fitting a linear model after thunderstorm cloud clustering with the improved GS-DBSCAN. The clustering analysis of lightning activity is carried out for canyons, ridges, valleys and river terrain, respectively. Finally, data from the South China Sea are selected to test the effectiveness of this paper's method in predicting lightning activity.

2. Recognition of lightning activity distribution patterns at different altitudes and latitudes and longitudes

The study of lightning activity law and its parameters is the basic work of lightning protection engineering design and risk assessment, and its related research has been the focus of extensive attention of scholars at home and abroad. Lightning activity has a strong geographical relationship and is affected by the topography of the study of lightning activity characteristics in different areas, not only for us to grasp and study the law of lightning activity in the place has a positive significance, but also has a positive guiding role in lightning disaster prevention and mitigation work.

Therefore, this section firstly takes the historical data of lightning monitoring network from 2015 to 2024 in Area C as the basis, and supplements the elevation information of the ground flash point based on the software, and systematically researches to obtain the changing rules of lightning key parameters such as lightning current amplitude, ground flash density, positive flash proportion, high (low) amplitude flash proportion with elevation, geographic latitude and longitude, and to identify the relevant lightning activity patterns.

2.1 Recognition of lightning activity patterns at different altitudes

Elevation is an important factor affecting lightning current magnitude. Given the large latitudinal span of the study area in this section, latitudinal corrections to the lightning current magnitude are required to obtain a more accurate model of the elevation influence relationship. Considering that the study area is mainly concentrated near the northern latitude, the following model is used to correct the lightning current magnitude at different latitudes:

$$I_{30} = I \times 10^{-10.66\times10^{-3}(30-\lambda)} \tag{1}$$

where I is the lightning current amplitude (kA) at the ground flash point before correction, λ is the degree of Han at the ground flash point, and I_{30} is the lightning current amplitude (kA) after correction to 30° latitude.

The study area in this section spans about 4.5° in longitude and 5° in latitude, with an elevation range of more than +90 m to more than +2800 m. The western area is mainly a plain with lower elevation, while the higher elevation alpine areas are mainly distributed in the northeast of the study area. The main lightning activities were concentrated in the altitude of 100 m to 1900 m, accounting for 98.55% of all lightning activities, and there were fewer lightning activities above 2000 m, only more than 1000, and the altitude distribution of lightning activities is shown in Fig. 1.

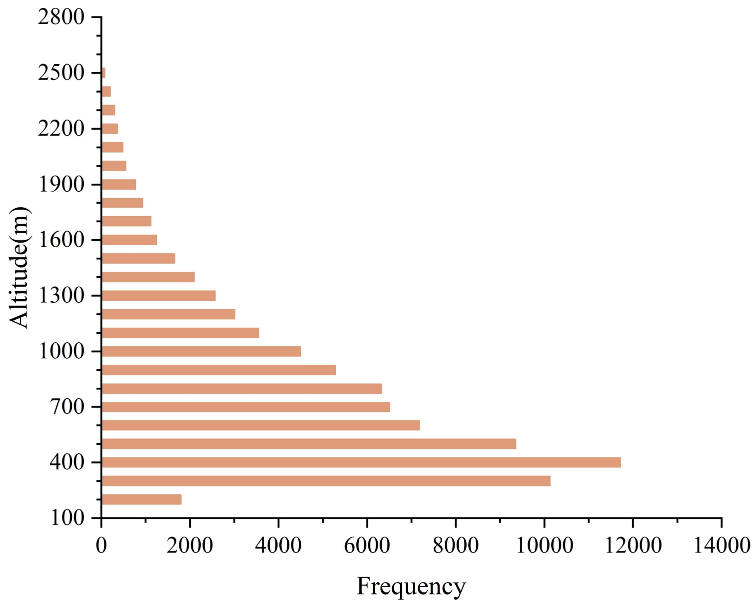


Fig. 1. Distribution of lightning interaction

2.2 *Recognition of lightning activity patterns at different latitudes*

Latitude is an important factor affecting the lightning current magnitude. In view of the large span of altitude in the study area, an altitude correction of the lightning current magnitude is required to obtain a more accurate relationship between the effects of latitude. For the convenience of the study, the following model is used to correct the lightning current magnitude at different altitudes:

$$I_{1000} = I \times 10^{-9.2\times10^{-5}(1000-H)} \tag{2}$$

where I is the lightning current amplitude (kA) at the ground flash point before correction, H is the elevation of the ground flash point, and I_{1000} is the lightning current amplitude (kA) after correction to an elevation of 1000 meters. The latitudinal span of the study area in this section is about 5°, and the elevation range is from more than 100 m to more than 2600 m. The western area is mainly a plain area with lower latitude, while the high mountainous area with higher latitude is mainly distributed in the northeast of the study area. The latitudinal distribution of lightning activity is shown in Figure 2, with the main lightning activity centered between 29° and 32° N latitude.

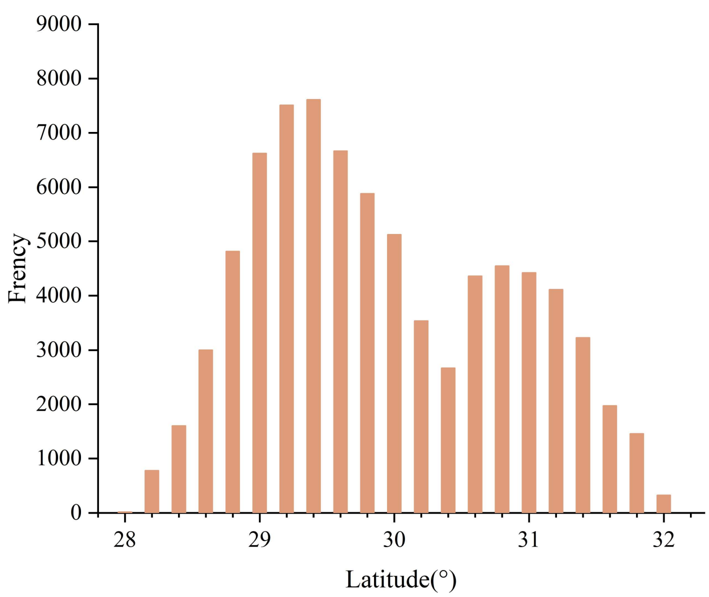


Fig 2. Latitude distribution of lightning activity

2.3 Recognition of lightning activity patterns at different longitudes

Longitude and latitude are important factors affecting the lightning current magnitude. Due to the large span of altitude in the study area, it is necessary to correct the lightning current amplitude for altitude in order to obtain a more accurate relationship between longitude and latitude. For the convenience of the study, the following formula is used to correct the lightning current magnitude at different altitudes:

$$I_{1000} = I \times 10^{-9.2\times10^{-5}(1000-H)} \tag{3}$$

Where I is the lightning current amplitude (kA) at the ground flash point before correction, H is the elevation of the ground flash point, and I_{1000} is the lightning current amplitude (kA) after correction to an elevation of 1000 meters. The frequency distribution of lightning activity at different longitudes is shown in Fig. 3. It can be seen that lightning activity is more frequent at 105.5°~106.8°E and 107.8°~109°E.

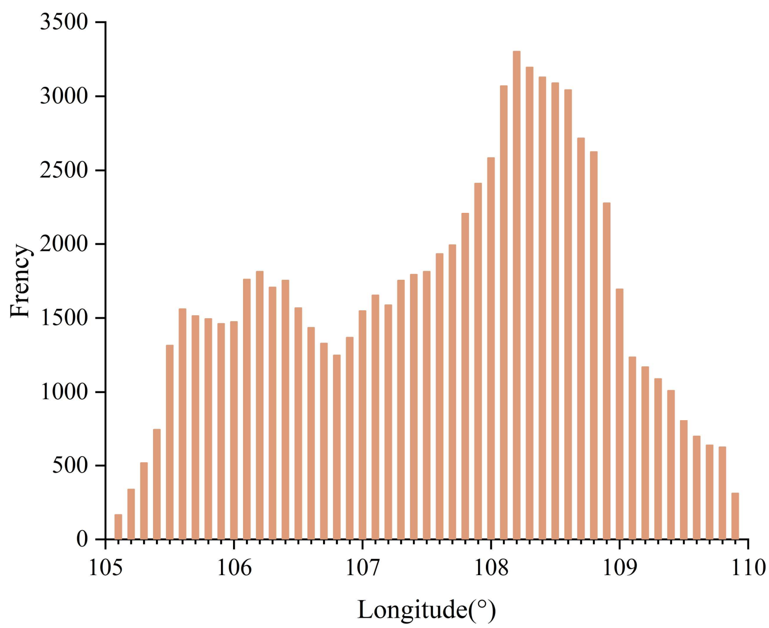


Fig. 3. Longitude distribution of lightning activity

3. Spatial prediction method for lightning activity based on GS-DBSCAN clustering

After completing the basic lightning activity pattern recognition, this paper improves the DBSCAN clustering method for spatial prediction of lightning activity.

3.1 Data pre-processing

Meteorological knowledge shows that thunderstorms are generally generated by thunderstorm cloud clusters, and thunderstorm cloud clusters have a longer duration, their horizontal coverage ranges from a dozen kilometers to hundreds of kilometers, and their cores are relatively stable, so the center of the cluster can be used to represent the center of the thunderstorm cloud cluster. The center of the thunderstorm cluster can be calculated using equation (4):

$$\begin{cases} X = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ Y = \dfrac{1}{n}\sum_{i=1}^{n} y_i \end{cases} \tag{4}$$

Where: X, Y are the latitude and longitude of the center of the cluster, respectively. x_i, y_i are the latitude and longitude of each core point in the cluster, respectively. n is the number of core points contained in each cluster.

3.2 Improved GS-DBSCAN clustering algorithm

3.2.1. Definition of clustering. Clustering is the process of dividing data elements into meaningful multiple clusters such that there is a large similarity between elements in the same cluster and a large dissimilarity between elements in different clusters. Since the division strategy of clustering is not unique, there exist a variety of clustering algorithms. The following five major categories are common:
1) Division-based clustering: first determine how many classes the objects in the database are clustered into, and use some kind of iterative reset technique to improve the division by moving the objects between divisions to divide the dataset into a pre-specified number of clusters of numbers.
2) Hierarchical based clustering methods: there are two main types, namely cohesive and split hierarchical clustering.

Cohesive hierarchical clustering is bottom-up, where each object in the database is treated as a clustered cluster, and similar clusters are merged successively until they are merged into a single cluster or a set termination condition is reached. Split hierarchical clustering is top-down, where clusters are decomposed according to the principle of maximum dissimilarity until there is only one object in each cluster or a set termination condition is reached.
3) Density-based clustering method: this algorithm divides the objects in the database into clusters by analyzing the distribution of the objects in the database and using the density as the similarity.
4) Grid-based clustering method: the basic principle is to divide the entire data space into a finite number of grid cells, map the set of data objects to the grid cells, and perform clustering on the grid.
5) Model-based clustering methods: each data cluster assumes a model before clustering and then finds the best fit of the data to the specified model.

3.2.2. DBSCAN algorithm DBSCAN is a classical density-based clustering method that randomly selects any point p from the dataset and finds all the points about Eps and $MinPts$ that are reachable from p density. According to the criterion that the number of points contained in a certain neighborhood is not less than $MinPts$, the region with sufficient density is divided into clusters corresponding to the largest set of points with density connected points.

The basic principle of the DBSCAN algorithm is: for the entire dataset, a point p is arbitrarily selected from the dataset and it is determined whether the point is a core point or not. If it is a core point, then find all the density reachable points starting from point p, and these points generate a cluster. Then the points in the cluster are added to the seed table, and the density reachable points from the seed point are found to extend the cluster, until no point is added, then a complete cluster is formed. If p is not a core point, label p temporarily as a noise point. Repeat the above steps for the next unprocessed point in the dataset for the next cluster extension. The clustering ends when all the points in the dataset have been processed [16].

The brief steps of the algorithm are as follows:
1) Input dataset with neighborhood radius Eps and threshold $MinPts$.
2) Pick any point p in the dataset and perform a region query.
3) If p is a core point, find all points that are density accessible from p and get a class containing p.

4) Otherwise, p is temporarily labeled as a noise point.
5) Examine the next point in the dataset and repeat the above steps until all points in the dataset are labeled as self-processed.

3.2.3. Geodesic distance. In this paper, the geodesic distance is chosen to replace the Euclidean distance. The geodesic distance is used to characterize the shortest distance between two given points on a connected surface, which represents the true distance between samples on a manifold structure and can reflect the global consistency of the data distribution [17]. In the study of nonlinear dimensionality reduction, a method of calculating the geodesic distance is generally used in the following steps:

Step 1 Input m dimensional dataset $D=(x_1,x_2,\cdots,x_n)$, n is the number of sample points. Calculate the k nearest neighbors of x_i based on the Euclidean distance.

Step 2 Initialize the geodesic distance:

$$d_G(x_i,x_j)=\begin{cases} d(x_i,x_j) & \text{If } x_j \text{ is a k-nearest neighbor point of } x_i \\ \infty & \text{Other} \end{cases} \tag{5}$$

$$d_{G_2}(x_i,x_j)=\min\left(d_G(x_i,x_j),d_{G_1}(x_j,x_i)\right) \tag{6}$$

Where $d(x_i,x_j)$ represents the Euclidean distance between point x_i and point x_j.

Step 3 calculates the shortest distance between two points of the data set:

$$d_G(x_i,x_j)=\min_{h=1,2,\cdots,n}\left\{d_{G_2}(x_i,x_j),d_{G_2}(x_i,x_h)+d_{G_2}(x_h,x_j)\right\} \tag{7}$$

3.2.4. Shared Nearest Neighbor Similarity. Considering the influence of the region around the data points, the similarity SNN is introduced by the number of nearest neighbors shared between the data points. SNN similarity is robust, it characterizes the local structure of the points in the data space, handles high dimensional data well, and is not very sensitive to changes in the density and dimensionality of the space [18].

The shared proximity of any two points x_i and x_j in the m -dimensional dataset $D=(x_1,x_2,\cdots,x_n)$ is:

$$SNN_r(x_i,x_j)=\left|n_r(x_i)\cap n_r(x_j)\right| \tag{8}$$

where $n_r(x_i)$ denotes the set of r nearest neighbors of x_i, i.e., the set consisting of the first r points closest to x_i. $|\cdot|$ denotes the number of elements contained in its set.

3.2.5. Similarity metrics based on geodetic distance and density. In this paper, a new similarity matrix is constructed to represent the distance metric between data points by geodesic distance, which overcomes the limitation of Euclidean distance. The shared nearest neighbor number SNN is then used to denote its density, which in turn is combined to obtain the similarity matrix between individual data points $W=\left[w_{ij}\right]_{n\times n}$. w_{ij} is expressed as follows:

$$w_{ij}=\begin{cases} \exp\left(-\dfrac{d_G^2(x_i,x_j)}{\sigma_i\sigma_j\left(SNN_r(x_i,x_j)+1\right)}\right) & i\neq j \\ 1 & i=j \end{cases} \tag{9}$$

where σ_i denotes the Euclidean distance from point x_i to the l rd nearest neighbor point.

3.2.6. Selection of parameters Eps and MinPts. Traditional DBSCAN has some difficulties in taking the values of Eps and MinPts, and the clustering results are sensitive to the parameters. In this paper, we adaptively determine the parameters according to the distribution characteristics of the data. Calculate the geodetic distance from each point in D to other points to get the distance matrix $G_{n\times n}$:

$$G_{n\times n}=\left\{d_G\left(x_i,x_j\right)|1\leq i,j\leq n\right\} \tag{10}$$

The $G_{n\times n}$ matrix is a real symmetric matrix with 0 diagonal elements. For each column element of $G_{n\times n}$ sort them in increasing order, and the elements of row $k+1$ are the distances from each object to the k th nearest neighbor.

3.2.7. GS-DBSCAN algorithm. Step 1 Input the original dataset S with parameter Eps and MinPts values. And the data set S is $z-score$ standardized to transform the data of different magnitudes into $z-score$ scores of uniform measure, and the preprocessed data set D is obtained.

Step 2 Calculate the geodetic distance $d_G\left(x_i,x_j\right)$ between data points according to Eq. (7), and calculate the $SNN_r\left(x_i,x_j\right)$ of any two points x_i and x_j in the dataset by Eq. (8), and the r value is taken as 5% of the number of sample points in the dataset. Then the similarity w_{ij} between the sample points based on the geodetic distance and SNN is calculated according to equation (9).

Step 3 After obtaining the similarity matrix W, select any point p in D whose Eps neighborhood is the set of sample points contained in a hypersphere region centered at p and with radius Eps is satisfied:

$$N_{Eps}(p)=\left\{q\in D\mid w_{pq}>\varepsilon\right\} \tag{11}$$

Step 4 Determine whether p is a core point, get the number of sample points $\left|N_{Eps}(p)\right|$ in the Eps neighborhood of p. If $\left|N_{Eps}(p)\right|$ is greater than or equal to the specified MinPts value, then p is a core point. Find all the points that are densely reachable from p to form a class C, and label p as processed. Otherwise, p is temporarily labeled as a noise point.

Step 5 After determining class C, repeat step 4 to process all the points in C to extend C.

Step 6 visits data points beyond C until all points in the dataset have been examined and clustering results are obtained.

3.3 Lightning activity prediction algorithm and process

3.3.1. Kriging Interpolation. According to the principle of Kriging interpolation, the general formula is as follows [19]:

$$Z\left(x_0\right)=\sum_{i,1}^{n}\lambda_i Z\left(x_i\right)(i=1,2,\cdots,n) \tag{12}$$

Where: $Z\left(x_0\right)$ is the predicted value at x_0, $Z\left(x_i\right)$ is the true value at x_i, and λ_i is the weight coefficient to be sought.

Assuming that $Z(x)$ satisfies the second-order smoothness condition throughout the sample space, based on the unbiasedness requirement: $E\left[Z^*\left(x_0\right)\right]=E\left[Z\left(x_0\right)\right]$, which is derived from equation (12):

$$\sum_{i=1}^{n}\lambda_i=1 \tag{13}$$

Minimize the estimation variance under unbiased conditions, i.e:

$$Min\left\{Var\left[Z^*\left(x_0\right)-Z\left(x_0\right)\right]-2\mu\sum_{i=1}^{n}\left(\lambda_i-1\right)\right\} \tag{14}$$

where μ is the Lagrange multiplier.

The matrix form when solving for the weight coefficients λ_i expressed in terms of the variational function $\gamma(x_i, x_j)$ can be obtained as:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ 0 \end{bmatrix} + \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n} \\ 1 \end{bmatrix} \tag{15}$$

The inverse of the matrix yields the weighting coefficients $\lambda_i (i = 1, 2, \cdots, n)$, from which the estimate at sampling point x_0 can be obtained $Z^*(x_0)$. where the variational function is:

$$\gamma(x_i, x_j) = \gamma(x_i - x_j) = E\left[Z(x_i) - Z(x_j)\right]^2 / 2 \tag{16}$$

3.3.2. Least Squares. Usually, when we study the interrelationships between variables (x, y), we start with a series of data $(x_i, y_i)(i = 1, 2, \cdots, m)$ in groups, plot these points in a $x - y$ coordinate system, and then use a curve to represent the relationship between these variables, which is a curve fitting problem. Least squares is one of the more common curve fitting methods, which applies the principle of squaring and minimizing the error in the fitting process to match the best function of the data [20].

Assuming that the data is $(x_i, y_i)(I = 1, 2, \cdots, m)$ and the expression is $y = f(x, a_0, \cdots, a_n)$, the sum of squares of error expression is:

$$F(a_0, \cdots, a_n) = \sum_{i=1}^{m} \left[f(x_i, a_0, \cdots, a_n) - y_i \right]^2 \tag{17}$$

Where: $a_0, \cdots, a_n$ is the pending parameter of the curve and m is the number of given points. The pending parameter can be obtained by solving the system of derivative equations, which is necessary to obtain the extreme values:

$$\frac{\partial F}{\partial a_k} = 0, (k = 0, 1, \cdots, n) \tag{18}$$

Based on Eq. (18), $n + 1$ equation can be obtained, and the parameters to be determined for curve $y = f(x, a_0, \cdots, a_n)$ can be derived immediately. In this paper, variable (x, y) is the latitude and longitude of the thunderstorm cloud center, and the curve obtained is the thunderstorm cloud trajectory.

3.3.3. Prediction Algorithm Flow. The improved DBSCAN clustering algorithm, which can adaptively determine the clustering parameters, and the weighted Euclidean distance is conducive to the elimination of noise points and the retention of core points, is more suitable for the identification and clustering of thunderstorm clouds, and lays a good foundation for further thunderstorm prediction.

The specific steps of thunderstorm warning are as follows:

1) Apply trans-Gaussian kernel density estimation to determine the parameters ε and MinPts for the selected ground flash data set D.
2) Select a point as the starting point, mark p it, and further search the $\varepsilon -$ neighborhood of the point via the weighted Euclidean distance formula to determine whether it is a core point.
3) If the point is a core point, create a new clustering cluster C and find out all the points in the neighborhood that are directly density accessible to join the candidate set N.
4) Judge whether all points in data set N are traversed, if not, repeat steps (2)-(4).

5) Combine all density reachable points to extend cluster C.
6) Output the set of target clusters to generate thunderstorm clouds.
7) Find the center of each cluster, and use kriging interpolation to determine the thunderstorm area.
8) Apply least squares fitting to the center of thunderstorms for multiple time periods, and generate the location of the lightning activity track.
9) Calculate the direction and speed of thunderstorm movement, and predict the location of thunderstorms and the area of lightning fall in the next time period.

4. Example analysis of lightning activity distribution pattern recognition and prediction

4.1 Cluster analysis of lightning ground flash activity in multiple terrains

In order to quantitatively represent the influence of topographic lightning activity, the variable overlap degree is defined in the paper to characterize the influence of topography on lightning activity. Let the number of lightning activities in the region in a certain period of time be m, get m lightning trajectories, calculate the n lightning trajectories that overlap with the terrain, and the degree of coincidence is n/m. Extract the contour lines from the topographic map, and get a number of lines with different terrain directions, and in this paper, we take three of them as an example, which are labeled as lines a, b, and c, respectively.

4.1.1. Canyon windway topography. Figure 4 shows a schematic diagram of typical lightning clustering activities in the neighboring canyon windway terrain. Using the clustering method to count the movement trajectories of lightning activity in the areas selected in this paper, it can be obtained that, for all the lightning activity in the area adjacent to line a, there are 4 paths that overlap with the canyon trend, and the calculated degree of overlap is 35.1%. b For all the lightning activity in the area adjacent to line b, there are 15 paths that overlap with the canyon trend, and the calculated degree of overlap is 33.4%. The remaining paths were partially along the northwest-southeast alignment, with 11 paths and 22.5% of the lightning trajectories overlapping with the river alignment. This is due to the NW-SE alignment of the river near the canyon.

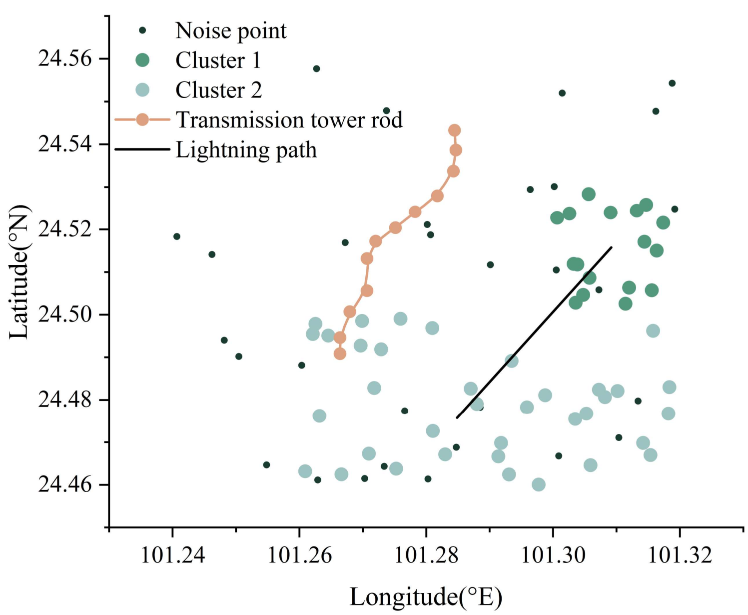


Fig 4. Typical lightning clustering activity on adjacent canyon wind passage terrain

4.1.2. Ridge topography. It can be seen from the statistics that the average overlap of the lightning activity trajectories along the ridge strike and slope of the ridges adjacent to the transmission lines is at 41.27%, except that the overlap of the lightning activity in the ridge terrain in the area adjacent to line a is 34.1%, and the overlap in the area adjacent to the other transmission lines does not differ much. Typical lightning clustering activities in the neighboring ridge terrain are shown in Fig. 5.

Typically, ridges are typical scenes that easily attract lightning because of their high terrain and prominent openings, and transmission towers located on the top of ridges are equivalent to another towering building in the towering terrain, which has a strong attraction capacity for lightning. Secondly, the transmission line passes through the top of the ridge, both sides of the ridge are slope surface, and the slope surface is terrain elevation type, the airflow is easy to rise along the slope, forming clouds at the top, resulting in a large number of falling mines on the slope surface as well.

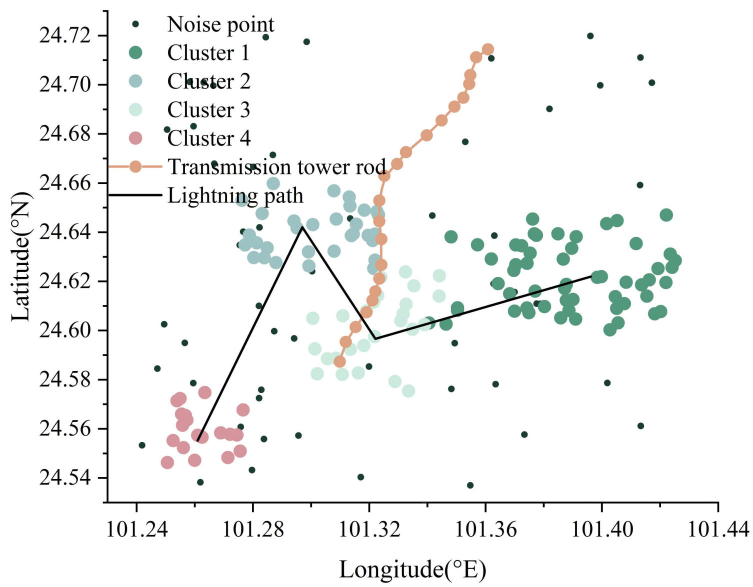


Fig. 5. Typical lightning clustering activity on adjacent ridge terrain

4.1.3. Valley terrain. Typical lightning clustering activity in near-valley terrain is shown in Figure 6. By counting, the average overlap of lightning activity trajectories along the valley alignment is at 35.64%. It can be seen that the surrounding area is open and located in a residential area, which makes it easy to attract lightning.

Usually, the valley has a low terrain with towering mountains on both sides, which attracts most of the lightning activity, and the flat terrain in the valley is narrow and stretched out, and influenced by the mountains on both sides, the airflow is easy to move along the direction of the valley from a fixed direction, but the direction of the airflow movement usually does not have any obvious obstacles to the movement of the airflow, or cause the airflow to rise dramatically to the convergence of the topography, and the low-altitude thunderclouds are pushed by the wind to move along the mountains, so some of the lightning activity will move in the direction of valleys, but the amount is much less than the lightning activity in ridge terrain.

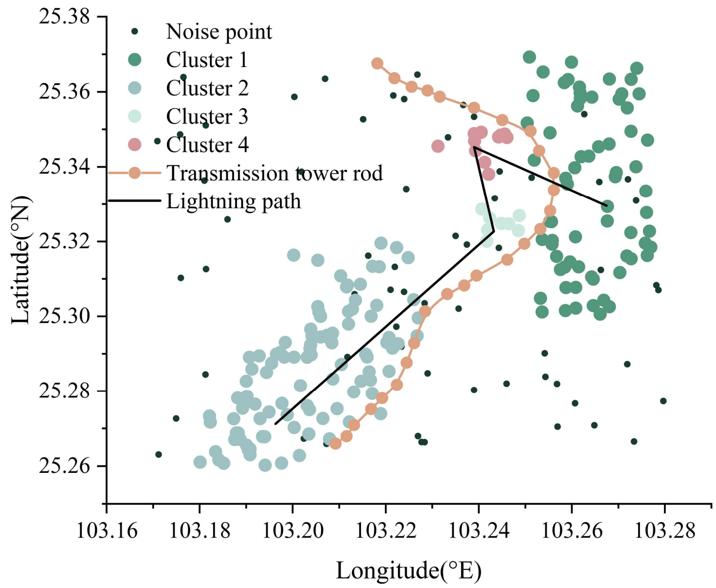


Fig. 6. Typical lightning clustering activity on adjacent valley terrain

4.1.4. Riverine topography. A schematic diagram of typical lightning clustering activities in neighboring river terrain is shown in Fig. 7. Through statistics, the average overlap of lightning activity trajectories along the river course is at 35.6%, and the overlap of lightning activity in the water adjacent to line c is much lower than the average because the river is fine and the water area is not large, which does not have a significant impact on the lightning activity.

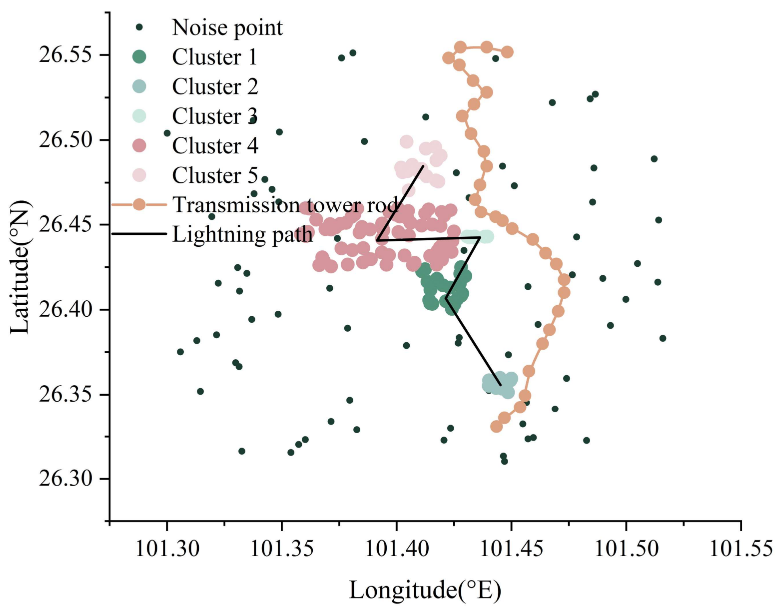


Fig. 7. Typical lightning clustering activity on adjacent river terrain

4.2 Simulation demonstration and analysis of lightning activity prediction

4.2.1. Justification of results. Python3.10 is used as the platform for program simulation, and ArGIS10.2 is used as the development tool. The original data were obtained from the thunderstorm data in southeastern China and the adjacent waters of the South China Sea monitored by the lightning detection and localization system. Due to the large amount of data, the thundercloud center-of-mass movement data of the thunderstorm data in the adjacent waters of the South China Sea in 2022-2023 were intercepted as historical data for the least-squares fitting of the linear equations. E127.5°, N23.0°-N27.0° range on August 1, 2022 to validate the prediction with data from a strong thunderstorm that occurred within the range of E127.5°, N23.0°-N27.0°.

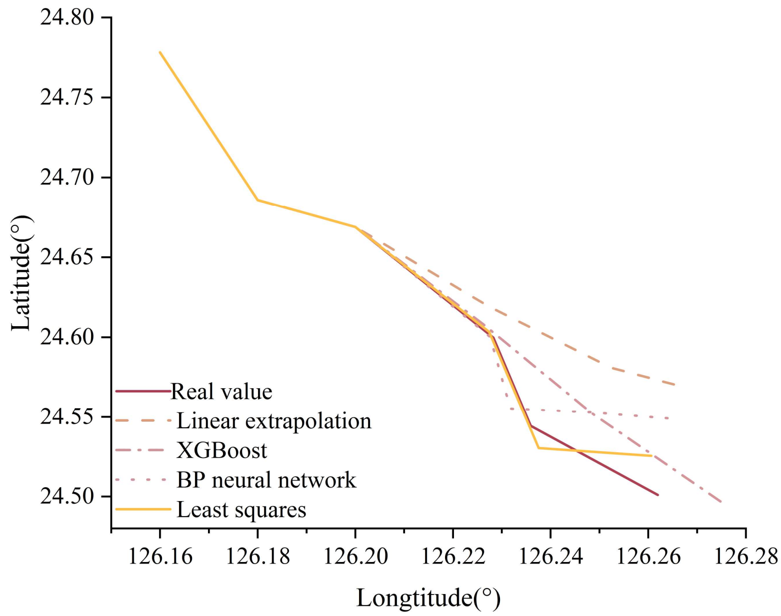
Firstly, the lightning localization data between 8:05 and 8:20 are divided into three segments according to the 5-min intervals, which are denoted by T1, T2, and T3, respectively, and then clustered by using the improved DBSCAN to generate the neighborhood radius and density thresholds based on the distribution characteristics of the lightning localization datasets of the three time segments themselves.

The center of mass of the thundercloud is used to approximate the replacement of the position of the entire clustered thundercloud, and the coordinates of the center of mass of the three thundercloud clusters are calculated for the time periods 8:05-8:10, 8:10-8:15, and 8:15-8:20, and the coordinates of the center of mass of the clustered cloud clusters for each time period are shown in Table 1. Then, the regression equation fitted by the historical thundercloud centroid movement trend data was used to predict the trajectory of the thundercloud centroid in the next 5 min, 10 min and 15 min time periods, and the longitude and latitude of the three thundercloud centroid in the three time periods of 8:05-8:10, 8:10-8:15 and 8:15-8:20 were taken as the input values, and the output values were the predicted longitude and latitude of the thundercloud centroid in the three consecutive time periods of 8:20-8:25, 8:25-8:30 and 8:30-8:35.

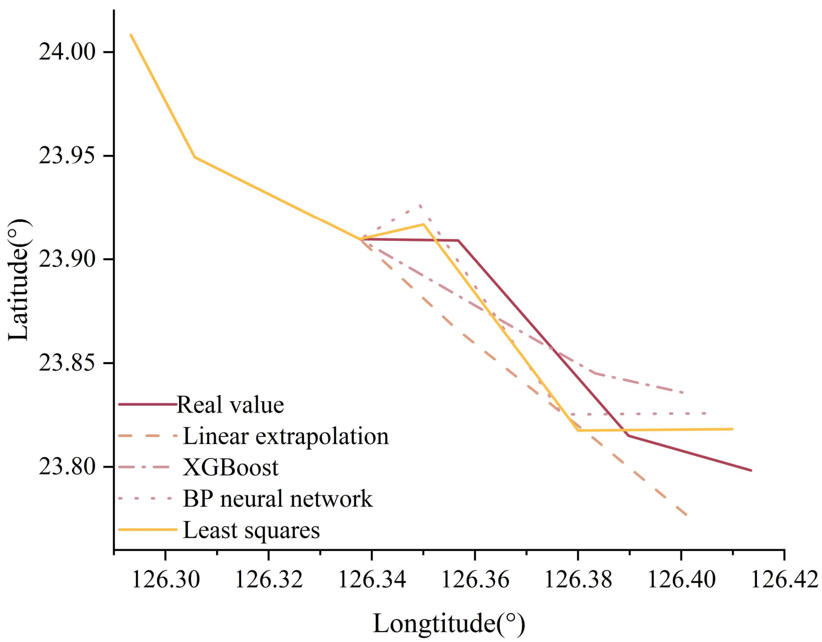
Table 1. Clustered thundercloud cluster center of mass calculation results

Time	Cluster cloud	Center of mass coordinates	
		Longitude°	Latitude°
8:05-8:10	C1	126.693	26.184
	C2	126.138	24.856
	C3	126.432	23.948
8:10-8:15	C1	126.774	26.161
	C2	126.179	24.780
	C3	126.318	23.852
8:15-8:20	C1	126.646	26.099
	C2	126.169	24.767
	C3	126.311	23.812

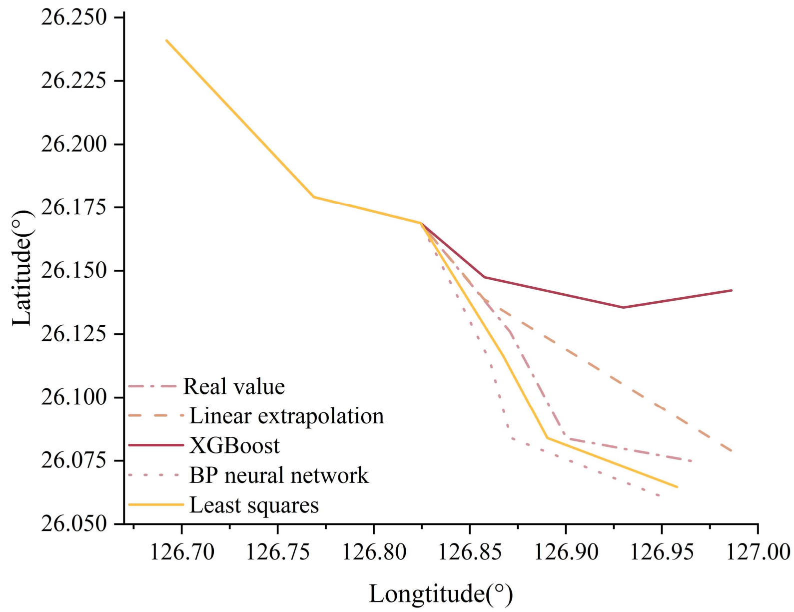
Linear extrapolation, XGBoost and BP neural network are introduced to compare with the least squares fitting algorithm, and the comparison results are shown in Fig. 8, (a) to (c) are the results of comparing the predicted values with the real values of the three clouds, respectively. The inputs of the first three times are the longitude and latitude of the thundercloud center of mass for the past 15 min, 10 min and 5 min, and the outputs of the last three times are the latitude and longitude of the thundercloud center of mass for the next 5 min, 10 min and 15 min. The initial thundercloud center of mass is located in the northwest direction, and the thundercloud center of mass has a tendency to move to the southeast direction as time passes. Compared to the other three algorithms, the predicted values of the least squares method are closer to the true values.



(a) The comparison of the predicted value of the cloud cluster c1



(b) The comparison of the predicted value of the cloud cluster c2



(c) The comparison of the predicted value of the cloud cluster c3

Fig. 8. Results of comparison between predictions of different algorithms and actual values

4.2.2. Error analysis. The difference between the latitude and longitude coordinates of the predicted center of mass and the actual center of mass is converted to the actual distance, and the actual offset distance of each prediction method is calculated as shown in Table 2, and the least-squares method has the smallest average offset error in the prediction of the center of mass of thunderclouds compared with the other three algorithms, and the average offset error is within 1.1 km, and the predicted values are closer to the real values.

Table 2. Prediction offset error by method

Cloud	Time	Deviation error(km)			
		Linear extrapolation	XGBoost	Least squares	BP neural network
C1	8:20-8:25	1.8	2.1	0.7	0.8
C1	8:25-8:30	3.5	4.3	0.9	1.8
C1	8:30-8:35	2.9	5.0	2.6	1.6
C2	8:20-8:25	0.4	1.4	0.3	1.0
C2	8:25-8:30	1.5	3.0	0.8	1.0
C2	8:30-8:35	0.9	3.7	1.7	1.2
C3	8:20-8:25	2.6	2.4	0.4	1.9
C3	8:25-8:30	0.9	1.9	0.9	1.3
C3	8:30-8:35	2.2	3.3	1.7	2.3
	Mean	1.9	3.0	1.1	1.4

In order to further the obtained lightning activity trajectory prediction results, false warning rate and false alarm rate are introduced, and the following concepts are related. Effective warning: an event for which a warning has been issued prior to the occurrence of lightning. This means that the impending lightning activity was successfully warned. Missed warning: refers to an event where no warning has been issued prior to the occurrence of lightning. It means that the lightning warning failed to recognize the impending lightning activity in time. False warning: an event where a warning is issued but lightning does not occur in the target area. This indicates a bias in the warning. False Alarm

Rate: Calculated by dividing the number of false warning events by the total number of valid warnings and false warning events; in short, the false alarm rate reflects the frequency of situations in which warnings are issued but lightning events do not actually occur.

The results of calculating the prediction accuracy for each time period are shown in Table 3. Due to the relative expansion of the thundercloud range after the interpolation of the inverse distance weights, the false alarm rate becomes an important indicator for assessing whether the predicted area of falling thunderstorms can truly and reasonably reflect the thundercloud mass, and due to the large prediction range, the spacing of the grid divisions is selected to be 0.05°. From Table 3, it can be seen that the prediction accuracy of this paper's prediction method is above 85%, and the false alarm rate (POD) is below 35%, which is a good prediction performance.

Table 3. Lightning prediction accuracy

Time	Predicted number of thunder/total thunder	Accuracy	Time	Wrong warning grid number/ total grid	POD
8:20-8:25	179/205	87.32%	8:20-8:25	61/213	28.64%
8:25-8:30	188/221	85.07%	8:25-8:30	72/211	34.12%
8:30-8:35	208/243	85.60%	8:30-8:35	75/221	33.94%

5. Conclusion

This study mainly proposes an improved GS-DBSCAN cluster analysis algorithm, combined with the least squares method to fit the prediction model, to realize the distributed pattern recognition and prediction of lightning activity. From the clustering analysis results, the spatial distribution of lightning activity overlaps with the canyon, ridge and valley strike by about 35%, reflecting the influence of topography on the distribution of lightning activity. Linear extrapolation, XGBoost and BP neural network are introduced to compare with the least squares fitting algorithm used in this paper. Comparing the real values with the predicted values, it is found that the prediction results of the method used in this paper are closer to the real values. The average offset error of the prediction is within 1.1 km, and the accuracy of the prediction method in this paper is higher than 85%, and the false alarm rate is lower than 35%. It shows that the method in this paper can effectively recognize the lightning activity distribution pattern and has high prediction accuracy.

Funding

This research was supported by the Inner Mongolia Autonomous Region Natural Science Foundation, Research on Cloud-to-Ground Lightning Characteristics Based on Spatiotemporal Clustering (2022MS04003); Hohhot Meteorological Bureau Innovation Project, Research on the Lightning Monitoring and Early Warning System in Hohhot City (hsqxkjzx202404).

References

[1] Bala, K., Choubey, D. K., Paul, S., & Lala, M. G. N. (2018). Classification techniques for thunderstorms and lightning prediction: A survey. In Soft-Computing-Based Nonlinear Control Systems Design (pp. 1-17). IGI global.

[2] Bao, R., Zhang, Y., Ma, B. J., Zhang, Z., & He, Z. (2022). An artificial neural network for lightning prediction based on atmospheric electric field observations. Remote Sensing, 14(17), 4131.

[3] Mansouri, E., Mostajabi, A., Tong, C., Rubinstein, M., & Rachidi, F. (2023). Lightning Nowcasting Using Solely Lightning Data. Atmosphere, 14(12), 1713.

[4] Rabbani, K. M. G., Islam, M. J., Fierro, A. O., Mansell, E. R., & Paul, P. (2022). Lightning forecasting in Bangladesh based on the lightning potential index and the electric potential. Atmospheric Research, 267, 105973.

[5] Geng, Y. A., Li, Q., Lin, T., Jiang, L., Xu, L., Zheng, D., ... & Zhang, Y. (2019, July). Lightnet: A dual spatiotemporal encoder network model for lightning prediction. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2439-2447).

[6] Bates, B. C., Dowdy, A. J., & Chandler, R. E. (2018). Lightning prediction for Australia using multivariate analyses of large-scale atmospheric variables. Journal of applied meteorology and climatology, 57(3), 525-534.

[7] Shi, M., Zhang, W., Fan, P., Chen, Q., Liu, Z., Li, Q., & Liu, X. (2022). Modelling deep convective activity using lightning clusters and machine learning. Int. J. Climatol, 42, 952-973.

[8] Etten-Bohm, M., Yang, J., Schumacher, C., & Jun, M. (2021). Evaluating the relationship between lightning and the large-scale environment and its use for lightning prediction in global climate models. Journal of Geophysical Research: Atmospheres, 126(5), e2020JD033990.

[9] RAHEEM, B., Ogbuju, E., & Oladipo, F. (2023). Development of a Lightning Prediction Model Using Machine Learning Algorithm: Survey. Journal of Applied Artificial Intelligence, 4(1), 45-56.

[10] Gharaylou, M., Farahani, M. M., Mahmoudian, A., & Hosseini, M. (2020). Prediction of lightning activity using WRF-ELEC model: Impact of initial and boundary conditions. Journal of Atmospheric and Solar-Terrestrial Physics, 210, 105438.

[11] Wang, X., Hu, K., Wu, Y., & Zhou, W. (2023). A Survey of Deep Learning-Based Lightning Prediction. Atmosphere, 14(11), 1698.

[12] Mostajabi, A., Finney, D. L., Rubinstein, M., & Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. Npj Climate and Atmospheric Science, 2(1), 41.

[13] Alves, E., Leal, A., Lopes, M., & Fonseca, A. (2021). Performance analysis among predictive models of lightning occurrence using artificial neural networks and SMOTE. IEEE Latin America Transactions, 19(5), 755-762.

[14] Shin, K., Kim, K., & Lee, G. (2024). Supervised Learning-Based Prediction of Lightning Probability in the Warm Season. Remote Sensing, 16(19), 3621.

[15] Brodehl, S., Müller, R., Schömer, E., Spichtinger, P., & Wand, M. (2022). End-to-End Prediction of Lightning Events from Geostationary Satellite Images. Remote Sensing, 14(15), 3760.

[16] Nian Zhou,Ping Jiang,Shiliang Jiang,Leshi Shu,Xiaoxian Ni & Linjun Zhong. (2024). An Identification and Localization Method for 3D Workpiece Welds Based on the DBSCAN Point Cloud Clustering Algorithm. Journal of Manufacturing and Materials Processing(6),287-287.

[17] Michele Bucelli,Francesco Regazzoni,Luca Dede' & Alfio Quarteroni. (2024). Robust Radial Basis Function Interpolation Based on Geodesic Distance for the Numerical Coupling of Multiphysics Problems. SIAM Journal on Scientific Computing(6),B981-B1002.

[18] Huang Ruohe,Xiao Ruliang,Zhu Weifu,Gong Ping,Chen Jinhui & Rida Imad. (2021). Towards an efficient real-time kernel function stream clustering method via shared nearest-neighbor density for the IIoT. Information Sciences364-378.

[19] Elias Bojago. (2024). Spatio-temporal rainfall variability and trends using a Kriging-interpolation and Innovative trend analysis approach: the case of Wolaita zone, south Ethiopia. Discover Sustainability(1),495-495.

[20] Jie Yuan,Hao Ma & Yan Wang. (2024). Dimensionality Reduction Nonlinear Partial Least Squares Method for Quality-Oriented Fault Detection. Mathematics(24),3911-3911.