

RESEARCH ARTICLE

Regionalization and diagnosis of coastal sea fog in Jiangsu, China based on machine learning

Dawei Shi^{1, 2}, Chao Li^{2, 3, *}, Jiaren Yan¹, Runqing Lv³, Lei Hui¹, Yuxiang Ji¹

¹Lianyungang Meteorological Bureau, Lianyungang, Jiangsu, China. ²Key Laboratory of Traffic Meteorology, China Meteorological Administration, Nanjing, Jiangsu, China. ³Jiangsu Meteorological Observatory, Nanjing, Jiangsu, China.

Received: April 28, 2026; accepted: August 17, 2026.

Sea fog is a high-impact marine boundary layer weather phenomenon that poses severe threats to maritime transportation, coastal production, and marine ecosystems. Its accurate prediction remains a long-standing challenge in marine meteorology. The coastal area of Jiangsu Province, China spans a large north-south range with significant regional heterogeneity in sea fog occurrence, while traditional administrative division cannot support refined sea fog forecasting. This study aimed to establish an objective regionalization scheme for coastal sea fog in Jiangsu, China and develop interpretable regional sea fog diagnostic models. Based on hourly observation data from 6 coastal marine stations in Jiangsu during 2015 – 2016 provided by the State Oceanic Administration of China, the K-means clustering algorithm was adopted to objectively divide the coastal sea fog regions. The sea fog occurrence diagnostic models for each region were then constructed using the C4.5 decision tree algorithm. The results showed that the Jiangsu coast could be optimally divided into three continuous and independent sea fog regions from north to south with distinct diurnal and monthly variation characteristics of sea fog frequency in each region. Relative humidity was the most critical factor affecting sea fog occurrence in all three regions, while the secondary influencing factors differed across regions as wind direction for region I, sea surface temperature for region II, and seawater salinity together with sea surface temperature for region III. The C4.5 decision tree models achieved training accuracies above 85% and test accuracies above 80% for all three regions with satisfactory generalization performance. This study provided an objective regionalization basis and interpretable diagnostic tools for refined sea fog forecasting along the Jiangsu coast and offered technical support for coastal disaster prevention and mitigation.

Keywords: sea fog; regionalization; diagnosis; K-means algorithm; C4.5 algorithm.

*Corresponding author: Chao Li, Key Laboratory of Traffic Meteorology, China Meteorological Administration, Nanjing, Jiangsu 210008, China. Email: gwerlc@163.com.

Introduction

Sea fog refers to a catastrophic weather in which atmospheric water vapor condenses or congeals in the boundary layer of the sea, resulting in the atmospheric visibility to be less than 1 km [1]. The formation and dissipation of sea fog are

regulated by complex boundary layer dynamic processes including turbulent transport, radiative cooling, wind shear, and fog-top entrainment [2]. Meanwhile, microphysical processes such as aerosol activation, droplet condensation, and collision-coalescence also dominate fog evolution, and their

parameterization schemes are important sources of numerical forecast uncertainty [3]. Sea fog in China mainly occurs in the Yellow Sea and the north of the East China Sea [4]. Jiangsu Province is in eastern China, facing the Yellow Sea to the east. The Yellow Sea is one of the regions with the most frequent sea fog occurrence in China. Jiangsu belongs to the north-south climate transition zone with both northern and southern climatic characteristics and dual land-sea effects [5]. As the dominant cause of local low-visibility events, sea fog in coastal Jiangsu brings inconvenience to coastal residents' daily life, production, transportation, and maritime operations. Long duration sea fog with extremely low visibility often causes serious maritime traffic accidents. In addition, sea fog blocks solar radiation and deteriorates seawater quality, leading to mass mortality of mariculture shrimp and shellfish [6]. Overall, sea fog exerts a significant impact on maritime navigation, agricultural and fishery production, and socio-economic development in coastal areas.

It is difficult for a mesoscale numerical model to accurately reproduce the complex boundary layer processes in sea fog events, and there are still many bottlenecks in the numerical prediction of sea fog [7]. At present, traditional statistical methods based on historical marine meteorological data remain one of the most effective approaches for sea fog forecasting [8]. Early studies mostly adopted stepwise regression and other statistical analysis methods, and established sea fog forecast methods for Qingdao, Shandong, China and short-term objective forecast systems for the Yellow Sea and East China Sea [9]. In recent years, machine learning techniques have been gradually applied to sea fog research. Huang *et al.* compared multiple machine learning algorithms for sea fog visibility prediction in the Beibu Gulf and found that XGBoost achieved the optimal performance [10]. Zhang *et al.* also developed high-precision sea fog occurrence prediction models using machine learning methods [11, 12]. The generation, development, and dissipation of sea fog often show significant regional

characteristics. The frequency of sea fog in China increases with latitude, meaning that sea fog occurs more frequently in the north than in the south. The coastal areas of Jiangsu, China are in a sea fog prone zone. Accurate prediction of sea fog occurrence is of great practical significance for social security and economic development in coastal Jiangsu and even East Asia. The coastal area of Jiangsu, China has a large north-south span. Conventionally, it is divided into three coastal sections according to administrative jurisdictions, but such city-based regional division is not representative for sea fog generation and dissipation forecasting. Whether for refined sea fog weather forecast or for studying the occurrence and development of sea fog in different regions, it is necessary to objectively divide coastal areas according to the actual occurrence rules of sea fog. Therefore, exploring an objective regionalization scheme for coastal sea fog has important practical and scientific value. With the continuous advancement of big data, cloud computing, and artificial intelligence, machine learning technology has been widely used in many fields, and an increasing number of studies have applied it to meteorological research. Researchers developed decision tree-based models for road icing and heavy fog disaster diagnosis, which achieved relatively accurate prediction performance [13, 14]. Nevertheless, the application of machine learning to regional sea fog research remains limited.

K-means is one of the top ten classical clustering algorithms in data mining. As an unsupervised algorithm, K-means does not require prior knowledge and has been widely used in artificial intelligence, data mining, and pattern recognition. The C4.5 decision tree is a classical supervised classification and prediction algorithm and is improved based on the ID3 algorithm, which is originally designed for discrete attribute data. The C4.5 algorithm adds continuous attribute discretization based on ID3. This study used the K-means clustering algorithm to objectively delineate coastal sea fog regions in Jiangsu, China and employed the interpretable

C4.5 decision tree algorithm to construct concise and accurate sea fog diagnostic models for different regions. The results of this study provided technical support for refined sea fog forecast and disaster prevention in coastal Jiangsu, China.

Materials and methods

Data source

The hourly observation data used in this study were provided by the State Oceanic Administration of China (Beijing, China). The data covered six coastal stations in Jiangsu, China from 2015 to 2016 including Lianyungang Station (34.75°N, 119.45°E), Binhai Station (34.27°N, 120.10°E), Dafeng Port Station (33.23°N, 120.82°E), Waikejiao Station (32.77°N, 121.33°E), Yangkou Port Station (32.52°N, 120.97°E), and Lvsu Station (32.07°N, 121.62°E). The observed variables included 2-meter (2 m) height air temperature, sea-level pressure, relative humidity, 10-minute average wind direction, 10-minute average wind speed, precipitation, sea surface temperature, seawater salinity, and visibility. Compared with traditional manual observation data, this dataset had the advantages of high observation frequency and no subjective observation error. All data were processed by eliminating missing data in the dataset. Low visibility caused by precipitation during precipitation periods was not classified as sea fog and were also excluded. A sea fog event was strictly defined as visibility less than 1 km with no concurrent precipitation. After quality control, 63,410 valid data samples accounting for 60.3% of the total samples including 2,630 sea fog samples (4.1% of valid data) and 60,780 non-sea-fog samples (95.9% of valid data) were included in this study. All statistical analyses and algorithm implementations in this study were performed using Python 3.9 (Python Software Foundation, Wilmington, DE, USA) with pandas, numpy and scikit-learn libraries (<https://scikit-learn.org/>).

K-means clustering algorithm

The implementation process of the K-means algorithm in this study included the initiation of k clusters in the given dataset with the loss function defined as follows.

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

where J was the sum of squared errors within clusters (loss function). k was the number of clusters. C_i was the i -th cluster. x was a sample point in cluster C_i . μ_i was centroid of cluster C_i . $\|x - \mu_i\|$ was Euclidean distance between sample x and centroid μ_i . The centroid of each cluster was calculated as below.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (2)$$

where $|C_i|$ was the number of samples in cluster C_i . The algorithm randomly selected k samples as the initial centroid of each cluster, calculated the distance between all sample points and each cluster centroid, and assigned each sample to the nearest cluster. The centroids were updated iteratively until the loss function converges. The silhouette coefficient was used to determine the optimal number of clusters k . The silhouette coefficient combined intra-cluster cohesion and inter-cluster separation with a value range of $[-1, 1]$. A value closer to 1 indicated better clustering performance.

C4.5 decision tree algorithm

The algorithm selected classification attributes for each node based on the information gain ratio, split samples according to the attribute with the maximum gain ratio and continued recursive splitting until the stopping condition was met. Post-pruning was performed to eliminate leaf nodes with no significant contribution to the model. Assuming that S was a training set containing s data samples, and s_i was the number of samples belonging to class i , the probability that a sample belonged to class i was determined as follows.

$$p_i = \frac{s_i}{s} \quad (3)$$

where p_i was the probability of class i . s_i was the number of samples in class i . s was the total number of samples in the training set. The information entropy of the training set S was defined as below.

$$Info(S) = -\sum_{i=1}^n p_i \log_2(p_i) \quad (4)$$

where $Info(S)$ was information entropy of the training set. n was the number of categories. When the dataset was divided by attribute A , the information entropy of the partitioned subset was below.

$$Info_A(S) = \sum_{j=1}^v \frac{|S_j|}{|S|} Info(S_j) \quad (5)$$

where v was the number of values of attribute A . S_j was the subset of samples with the j -th value of attribute A . The information gain was then calculated as follows.

$$Gain(A) = Info(S) - Info_A(S) \quad (6)$$

where $Gain(A)$ was the information gain of attribute A . The gain ratio was then calculated as below.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (7)$$

where $SplitInfo(A)$ was the split information entropy of attribute A . One major improvement of the C4.5 algorithm over ID3 was its ability to process continuous data. The processing flow for continuous attributes was first sorting the attribute values in the node followed by dynamically dividing the training set with different thresholds. The midpoint of two adjacent values was taken as a candidate threshold before all samples being divided into two categories according to the threshold. The information gain and gain ratio for all candidate thresholds were then calculated followed by selecting the threshold with the highest gain ratio to discretize the continuous attribute into two branches.

Experimental design and evaluation metrics

The task of "whether visibility is less than 1 km" was abstracted as a binary classification task of "whether sea fog occurs". The hold-out method was adopted for model training and testing. The dataset was divided into mutually exclusive training set and test set. The training set was used to construct the decision tree model, while the test set was used to evaluate the generalization performance of the model. To minimize the impact on the original temporal distribution of data, an equidistant sampling strategy was used to divide the training and test sets. Samples were selected at 4-hour intervals from the three regions to form the test set (approximately 25% of the total data), and the remaining samples were used as the training set (approximately 75% of the total data). To address the sample imbalance between sea fog and non-sea-fog events, bootstrap sampling (sampling with replacement) was performed independently on sea fog samples in both the training set and test set, so that the number of sea fog samples was approximately equal to the number of non-sea-fog samples while retaining the original data characteristics. After balancing, the sample sizes of each region were determined as that 11,100 non-fog samples and 11,106 fog samples were in region I as the training set, while 3,705 non-fog samples and 3,705 fog samples were in the test set. For region II, 11,834 non-fog samples and 11,832 fog samples were in the training set, while 3,939 non-fog samples and 3,936 fog samples were in the test set. In region III, 22,660 non-fog samples and 22,473 fog samples were in the training set, while 7,542 non-fog samples and 7,554 fog samples were in the test set. A pre-pruning strategy was adopted to mitigate overfitting with the minimum number of samples at each leaf node being set to no less than 15% of the total samples. The model performance was evaluated by training accuracy, test accuracy, hit rate (probability of detection), and false alarm rate.

Results

Regionalization of coastal sea fog based on K-means clustering

The visibility data of six coastal stations from 2015 to 2016 were used to construct the K-means clustering model. The silhouette coefficients for k values from 1 to 5 were calculated and compared. The results showed that when $k = 3$, the silhouette coefficient reached the highest value, indicating the best clustering performance. Therefore, the number of clusters was set to 3. The objective regional division of coastal sea fog in Jiangsu, China was divided into three continuous regions from north to south as region I represented by Lianyungang Station and Binhai Station, region II represented by Dafeng Port Station and Waikexiao Station, and region III represented by Yangkou Port Station and Lvsu Station (Figure 1).

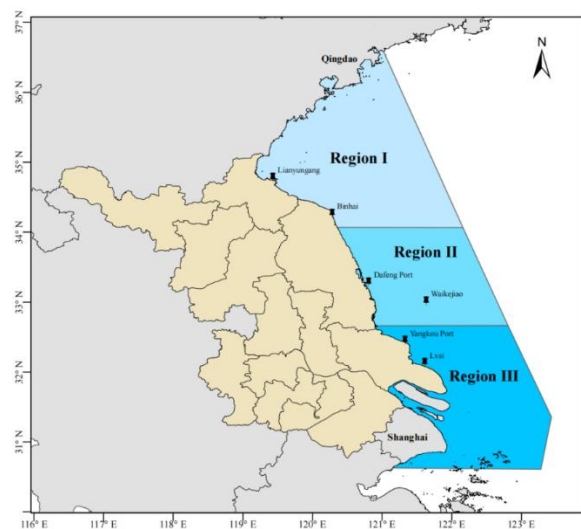


Figure 1. Division map of sea fog regions along the coast of Jiangsu, China based on K-means algorithm.

The diurnal distribution characteristics of sea fog frequency in different regions demonstrated that the peak and valley of sea fog frequency occurred at different times in the three regions. In region I, sea fog frequency peaked at around 7:00 and reached the minimum at 15:00 with the smallest diurnal fluctuation. In region II, the frequency peaked at around 6:00 and dropped to the lowest from 17:00 to 20:00 with a relatively large diurnal

fluctuation. In region III, the frequency peaked from 3:00 to 5:00 and bottomed out at around 15:00 with the largest diurnal fluctuation (Figure 2). Overall, the three regions demonstrated distinct diurnal patterns in both peak/valley timing and fluctuation amplitude.

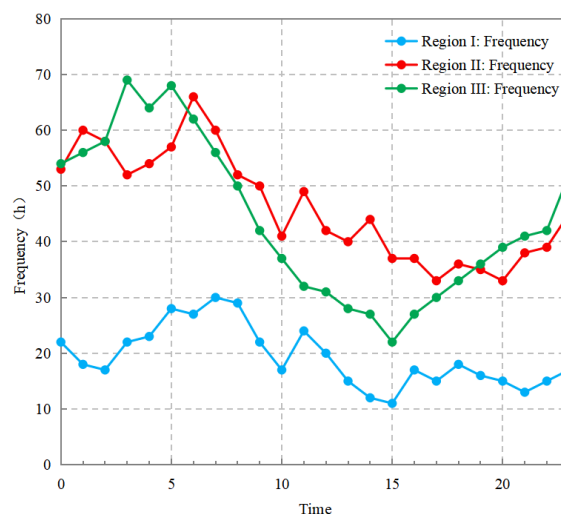


Figure 2. Diurnal distribution of sea fog frequency in different coastal regions of Jiangsu China from 2015 to 2016 during local time.

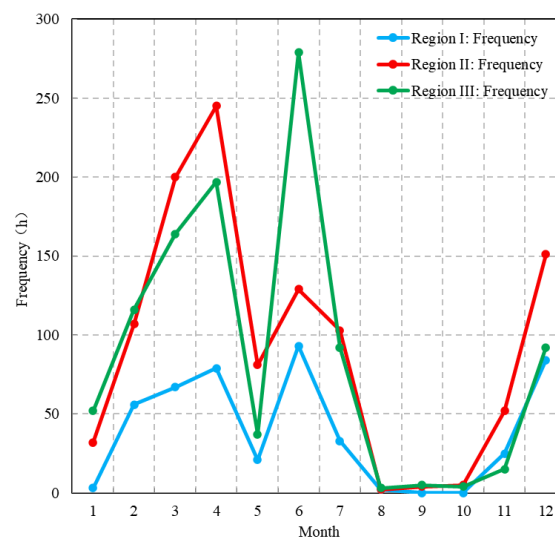
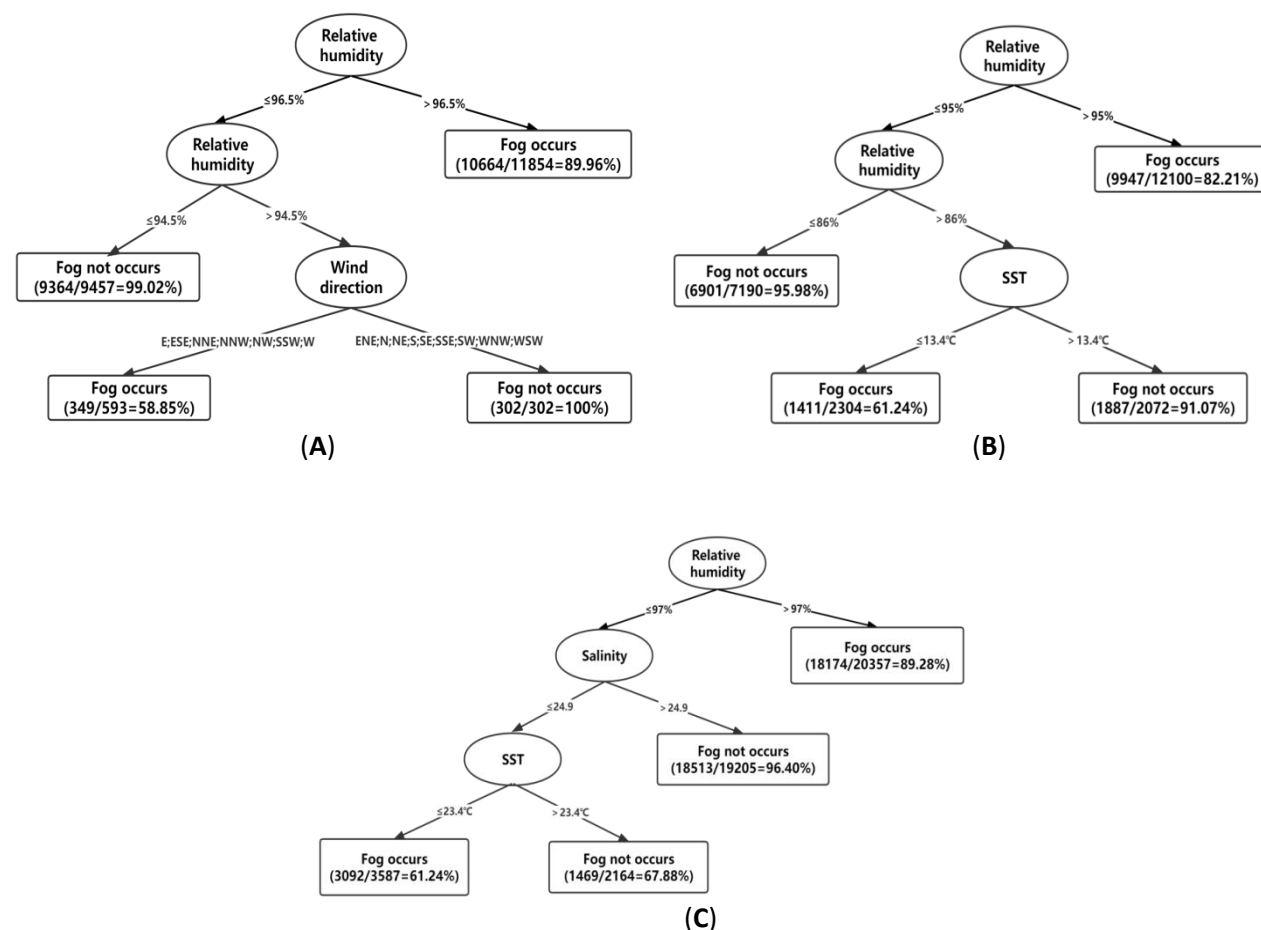


Figure 3. Monthly distribution of sea fog frequency in different coastal regions of Jiangsu, China from 2015 to 2016.

The monthly distribution of sea fog frequency in the three regions showed that in region I, sea fog

Table 1. Distribution of sea fog occurrence in different regions of Jiangsu, China.

Sea fog regions	Non-sea-fog frequency (h)	Sea fog frequency (h)	Sea fog occurrence proportion
Region I	14,805	463	3.03%
Region II	15,773	1111	6.58%
Region III	30,202	1056	3.38%

**Figure 4.** Decision tree for sea fog occurrence in region I (A), region II (B), region III (C). The values in parentheses indicated the number of correctly classified samples and the learning accuracy.

frequency was the highest in June and no sea fog occurred in September. In region II, the frequency peaked in April and was the lowest in August. In region III, the frequency reached the maximum in June and the minimum in August (Figure 3). In general, sea fog rarely occurred in all three regions from August to October. The results showed that the three regions divided by the K-means algorithm had independent spatial continuity and distinct sea fog occurrence

characteristics, which could well reflect the regional differences of sea fog development along the Jiangsu coast.

Sea fog diagnosis model based on C4.5 decision tree

The sample distributions of sea fog occurrence in the three regions were shown in Table 1, while the decision tree structures of the three regions were shown in Figure 4. For region I, relative

humidity was the root node, the most important factor. When relative humidity > 96.5%, sea fog was highly likely to occur. When relative humidity was between 94.5% and 96.5%, wind direction became the secondary classification factor with the sea fog more likely to occur under east, east-southeast, north-northeast, north-northwest, northwest, south-southwest, and west wind directions. For region II, relative humidity was also the root node. When relative humidity $\leq$ 86%, sea fog was unlikely to occur. When relative humidity was between 86% and 95%, sea surface temperature (SST) became the secondary classification factor with sea fog more likely to occur when SST $\leq$ 13.4°C. For region III, relative humidity was still the most critical factor, and seawater salinity and SST acted as secondary classification factors. In general, relative humidity was the most important factor for sea fog occurrence in all three regions, while the secondary important factors differed as wind direction for region I, SST for region II, and seawater salinity and SST for region III. The performance metrics of the three models showed that the training accuracy of region I was 93.12%, while region II and region III were 85.13% and 91.39%, respectively. The test accuracy of region I was 92.56%, while region II and region III were 84.29% and 90.59%, respectively. The hit rate (probability of detection) in regions I, II, and III were 98.22%, 94.31%, and 93.01%, respectively. The false alarm rate in region I was 13.10%, while regions II and III were 25.74% and 11.83%, respectively. The test accuracy referred to the overall classification correctness on the entire test set including both fog and non-fog samples, while the hit rate specifically denoted the proportion of actual sea fog events that were correctly identified. The false alarm rate was the proportion of predicted sea fog events that did not actually occur. Overall, the decision tree models showed good classification performance and robust generalization ability for all three coastal regions, providing a concise, interpretable, and valuable reference for sea fog forecasting along the Jiangsu coast.

Discussion

Sea fog is a complex marine boundary layer weather phenomenon. Although the decision tree algorithm can effectively explore the statistical rules of sea fog occurrence and development, it has limitations in explaining the underlying physical mechanisms. In-depth study of physical mechanisms is an important basis for statistical analysis. From the perspective of data applicability, the marine observation data from the State Oceanic Administration were more suitable for this study than conventional land meteorological data. The dataset contained both basic meteorological elements and marine observation elements such as sea surface temperature and seawater salinity, which could capture the influence of ocean underlying surface on sea fog. On the other hand, the observation stations were mostly located on islands or coastal shores, which were consistent with the actual occurrence environment of sea fog. The regional differences in key influencing factors revealed by the decision trees were potentially driven by different synoptic systems, ocean currents, and local geographical factors. For the northern part (region I), wind direction played a key role by regulating water vapor transport from the ocean. For the central part (region II), sea surface temperature dominated the air-sea temperature difference and thus affected condensation conditions. For the southern part (region III), the combined effect of seawater salinity and SST reflected the influence of more complex coastal hydrodynamic conditions. Further, the logical progression of this research was to evolve the current diagnostic model into a predictive one. Future work should focus on applying the established decision rules to numerical weather prediction forecast fields and validating the model's performance in a true forecasting context, which was a critical step toward operational application. With the continuous deepening of sea fog basic theory and the iterative upgrading of machine learning algorithms, the application of machine learning in sea fog diagnosis and prediction would achieve more meaningful results.

In this study, the K-means clustering algorithm was used to objectively regionalize coastal sea fog in Jiangsu Province of China, while C4.5 decision tree models for sea fog occurrence diagnosis were established for three regions. Based on visibility data from 6 marine observation stations, the Jiangsu coast could be optimally divided into three continuous and independent sea fog regions from north to south with distinct diurnal and monthly variation characteristics of sea fog frequency in each region. The results found that relative humidity was the most critical factor for sea fog occurrence in all three regions. The secondary influencing factors showed regional heterogeneity with wind direction for region I, sea surface temperature for region II, and seawater salinity combined with sea surface temperature for region III. The C4.5 decision tree models achieved satisfactory performance with training accuracies above 85% and test accuracies above 80% for all three regions, showing strong generalization ability and good application value for refined sea fog forecasting.

Acknowledgements

This work was supported by the Joint Fund of the National Natural Science Foundation of China (Grant No. U2542218), Jiangsu Young Scientific and Technological Talents Promotion Project (Grant No. TJ-2021-21), and the Open Fund of Nanjing Joint Institute for Atmospheric Sciences "Beijige" (Grant No. BJG202309).

References

1. Zhang P, Wang Y, Li J, Fang F, Zhu L, Lv J. 2025. Improved parameterization of cloud droplet spectral dispersion expected to reduce uncertainty in evaluating aerosol indirect effects. *Geophys Res Lett*. 52:e2024GL111643.
2. Shao N, Lu C, Li Y, Jia X, Wang Y, Yin Y, *et al*. 2025. Impact of improved surface flux parameterization on simulation of radiation fog formation in the Yangtze River Delta, China. *J Geophys Res Atmos*. 130(9):042345.
3. Wang Y, Xu X, Lu C, Zhu L, Wang X, Zhang P. 2025. Assessment of the corrected CCN activation parameterizations in simulating shallow cumulus using large-eddy simulations. *Atmos Res*. 315:107881.
4. Fu G, Song YJ. 2014. Climatology characteristics of sea fog frequency over the Northern Pacific. *J Meteorol Sci*. 34(1):35-41.
5. Chen XY, Zhang J, Ma Y. 2014. Analysis of the spatial and temporal changes of the coastline in Haizhou Bay during the past 40 years. *J Meteorol Sci*. 34(3):324-334.
6. Zhang SP, Bao XW. 2008. The main advances in sea fog research in China. *J Meteorol Sci*. 28(3):359-366.
7. Qu P, Wang W, Liu Z, Gong X, Shi C, Xu B. 2021. Assessment of a fusion sea surface temperature product for numerical weather predictions in China: A case study. *Atmosphere*. 12(5):604.
8. Jiang DS, Zhang SP, Lu WS. 2008. The climatic characteristics and forecast study of the sea fog in Qingdao. *J Meteorol Sci*. 28(3):7-12.
9. Yu RL, Wang YN, Li YP. 2007. Operational system for short-term objective sea fog forecasts in the Yellow Sea and the East China Sea. *Atmos Sci Res Appl*. 16(2):28-37.
10. Huang Q, Zeng P, Guo X, Lyu J. 2024. Utilizing machine learning and multi-station observations to investigate the visibility of sea fog in the Beibu Gulf. *Remote Sens*. 16(18):3392.
11. Zhang W, Leung Y, Chan JCL. 2013. The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I: Tropical cyclone recurvature. *J Appl Meteorol Climatol*. 52(6):1394-1416.
12. Zhang W, Leung Y, Chan JCL. 2013. The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part II: Tropical cyclone landfall. *J Appl Meteorol Climatol*. 52(6):1417-1432.
13. Shi DW, Geng HT, Ji C, Huang C. 2015. Construction and application of road icing forecast model based on C4.5 decision tree algorithm. *J Meteorol Sci*. 35(2):204-209.
14. Shi DW, Li C, Shi YM. 2018. Study on the localization diagnosis of extra heavy fog on the background of the fog weather based on machine learning algorithms. *J Catastrophol*. 33(2):193-199.
15. Quinlan J. 1987. Decision trees as probabilistic classifiers. *Proc 4th Int Workshop on Mach Learn*. Irvine, CA, USA. AAAI.