Meteorological and Environmental Research 2026, 17(4): 26 – 28

A Machine Learning-Based Lightning Intensity Prediction Model and Its Operational Forecasting Application

Shiyun LIU[1*], Nan ZHAO[2], Yanni SONG[2], Zhuojun ZHANG[3]

1. Inner Mongolia Meteorological Observatory, Hohhot 010051, China; 2. Ulanqab Meteorological Bureau, Ulanqab 012000, China; 3. Siziwang Banner Meteorological Bureau, Inner Mongolia Autonomous Region, Siziwang Banner 011899, China

Abstract Lightning activity is characterized by abrupt occurrence, strong spatial heterogeneity, and highly variable intensity, and traditional empirical methods are difficult to meet the needs of refined lightning warning. In this paper, the Ulanqab region was selected as the study area. VLF/LF lightning location data and ERA5 reanalysis data collected during the summers of 2020 – 2022 were used, and a lightning intensity level prediction model was constructed. Based on the environmental conditions of lightning occurrence, nine key predictors were selected from 25 environmental factors, including chimney index, total totals index, K index, 0 – 6 km of vertical wind shear, deep convection index, uplift condensation height temperature, dew point temperature, 700 hPa of pseudo equivalent temperature, and precipitable water vapor. Four machine learning methods, namely gradient boosting, random forest, neural network, and logistic regression, were used to establish a five-level lightning intensity classification model for weak, moderate, strong, very strong, and extreme lightning. The results showed that the gradient boosting model had the highest comprehensive score of 0.173 7, and performed the best in terms of average CSI, overall ACC, and comprehensive prediction ability. The DeLong test results indicated that the overall performance of gradient boosting was significantly better than other models ($P < 0.05$). Using the typical lightning process during August 18 – 20, 2024, a 72-hour historical simulation and EC business prediction verification were conducted. The results showed that the model can better reflect the changes in lightning activity intensity and spatial evolution patterns, and has the business forecasting application capability.

Key words Lightning intensity; Machine learning; Gradient boosting; Historical simulation; Operational forecasting

DOI 10.19547/j.issn2152 – 3940.2026.04.005

Lightning is one of the most important manifestations of severe convective weather, characterized by strong suddenness, small spatial scale, and high disaster risk[1]. With the rapid development of new energy, power facilities, and major engineering construction, increasing demands have been placed on refined monitoring and early warning of lightning activity[2–3]. Traditional lightning prediction mainly relies on weather situation analysis, empirical indicators, and physical parameter threshold judgment, which has certain limitations in complex nonlinear environments[4]. Machine learning methods can utilize a large amount of historical meteorological data to mine the nonlinear relationship between environmental variables and lightning activity, providing a new technological approach for objective lightning prediction[5–10]. This study focused on the characteristics of lightning activity in the Ulanqab region. Based on historical lightning location data and ERA5 reanalysis data, a multi model lightning intensity level prediction method was constructed. The optimal algorithm was determined through model performance comparison, and verification was conducted by combining ERA5 historical simulation and EC business forecast data, providing technical support for refined business prediction of lightning.

Received: June 11, 2026　　Accepted: July 30, 2026

Supported by the Guiding Innovation Fund of the Inner Mongolia Autonomous Region Meteorological Bureau (nmqxydcx202216).

* Corresponding author.

1 Data and methods

1.1 Data materials Lightning data were obtained from VLF/LF lightning location data from Ulanqab area, including information such as lightning occurrence time, latitude and longitude, polarity, and current intensity. The meteorological and environmental data used ERA5 hourly reanalysis data, including variables such as temperature, humidity, wind field, and geopotential height. During operational forecasting, EC numerical forecast data were used as future environmental input.

1.2 Classification of lightning intensity levels According to the local standard of Inner Mongolia (DB 15/T 2747 – 2022), the cloud-to-ground lightning intensity was divided into five levels: weak, moderate, strong, very strong, and extreme lightning using the percentile method. A classification prediction model was established between environmental parameters and lightning intensity level.

1.3 Screening of key convective parameters Based on the thermal, water vapor, and dynamic conditions of thunderstorm occurrence, Pearson and Spearman correlation analysis and feature importance evaluation were conducted on 25 environmental parameters. Finally, 9 key convective parameters were selected: chimney index (CHIMNEY), total totals index (TT), K index, deep convection index (DCI), lifting condensation level temperature (LCLT), dew point, 700 hPa of equivalent temperature, precipitable water (PW), and 0 – 6 km of vertical wind shear

(SHR06), covering three dimensions: thermal instability, water vapor conditions, and dynamic conditions.

1.4 Construction and evaluation of lightning intensity prediction model Four machine learning models were developed using random forest, gradient boosting (implemented using LightGBM), neural network, and logistic regression algorithms. The models were trained using time series cross validation, and hyperparameters were optimized using Bayesian optimization. The evaluation indicators included probability of detection (POD), false alarm rate (FAR), critical success index (CSI), accuracy rate (ACC), and overall business score. Among them, CSI was used to evaluate the consistency between predicted results and actual situations.

2 Model performance analysis

As shown in Table 1, the gradient boosting model had the best comprehensive prediction ability, with advantages in average CSI, overall ACC, and comprehensive business score. Different models had certain differences in lightning intensity level. But from a business application perspective, it was necessary to comprehensively consider the recognition ability, stability, and generalization ability of each level. Therefore, gradient boosting was ultimately determined as the business prediction model for lightning intensity. The DeLong test results showed that the overall performance of gradient boosting was significantly better than other models ($P < 0.05$), with only insignificant difference in the extreme lightning compared to some models. These results suggested that there was still improvement room for predicting extreme lightning events due to the small sample size. The analysis of feature importance showed that parameters such as CHIMNEY, LCLT, and SHR06 contributed significantly, indicating that thermal instability, water vapor supply, and dynamic uplift jointly affected changes in lightning intensity.

Table 1 Comprehensive evaluation results of four models

Model	Average CSI	Overall ACC	Business weighted rating	Overall rating
Gradient boosting	0.128 3	0.266 4	0.141 4	0.173 7
Random forest	0.127 2	0.257 5	0.141 5	0.170 6
Neural network	0.132 2	0.254 0	0.135 1	0.169 6
Logistic regression	0.102 0	0.193 6	0.107 2	0.131 1

3 Historical case simulation and business prediction verification

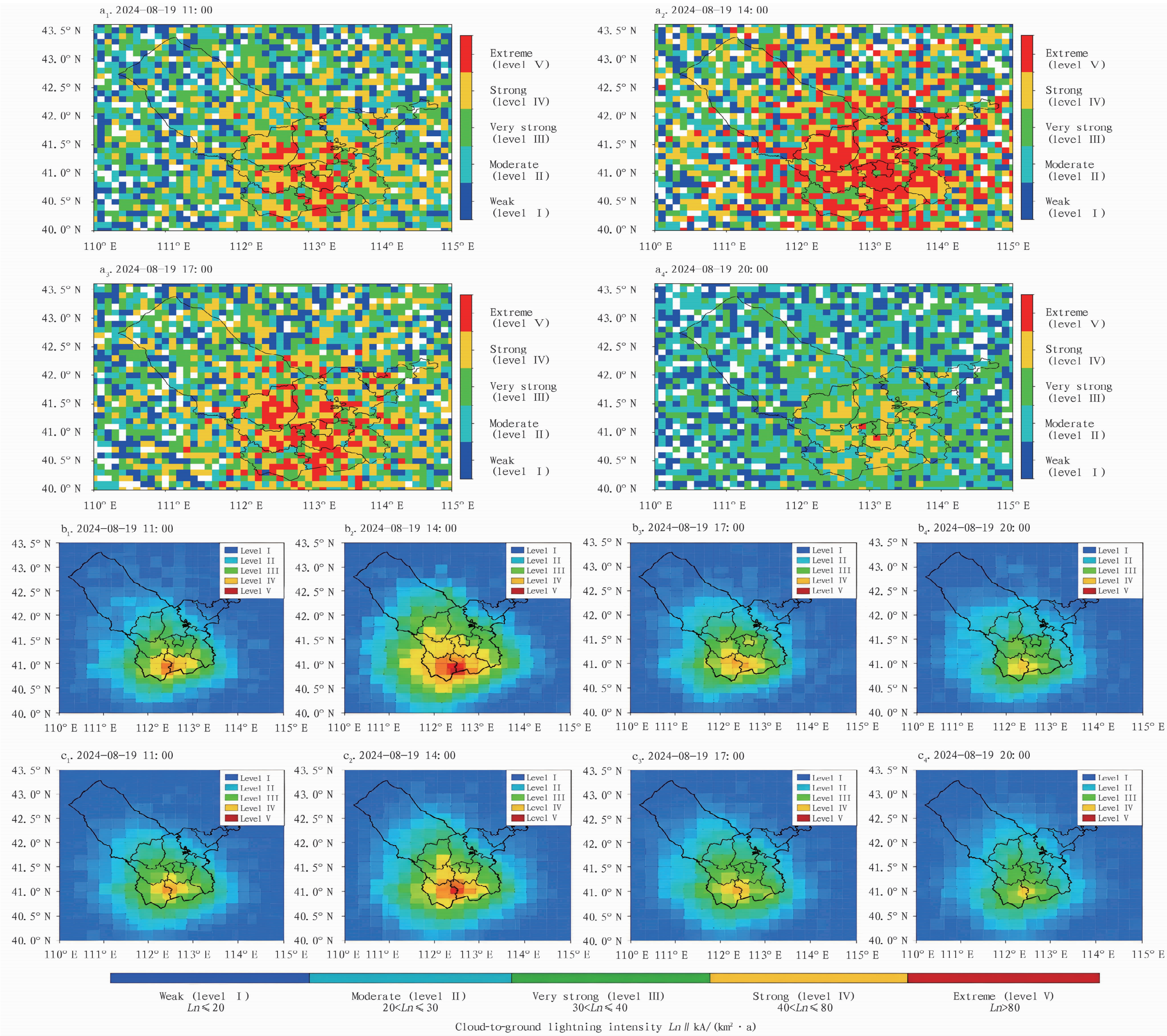
3.1 ERA5 historical simulation verification A typical lightning event occurring during 18 – 20 August 2024 was selected to conduct a 72-hour historical simulation. Based on ERA5 hourly environmental field data, inputting the trained gradient boosting model, the probability and intensity level of lightning occurrence were output. The results showed that lightning activity had obvious daily variation characteristics, mainly concentrated between 12:00 – 18:00, and reached its peak between 15:00 – 16:00. The intensity of lightning had a characteristic of periodic changes, with weak lightning intensity in the morning, the proportion of strong and very strong lightning increased in the afternoon, and gradually weakening in the evening. The occurrence probability of extreme lightning was lower, and the model can reflect the low-frequency characteristics of extreme lightning events.

3.2 EC business prediction verification EC numerical forecast data were used to replace ERA5 environmental field to conduct a 72-hour operational forecasting experiment. The results showed that in the early stage of the forecast, the risk of strong lightning was relatively high; the probability of stronger lightning increased in the middle stage; the probability of moderate and weak levels increased in the later stage, reflecting the gradual weakening process of the convective system. The probability of extreme lightning throughout the entire process remained at a low level, consistent with the actual characteristics of lightning activity.

3.3 Comparison between ERA5 simulation and EC prediction Both the ERA5 historical simulation and the EC operational forecast reproduced the center location, intensity trend, and temporal evolution of lightning activity. Their spatial distributions were generally consistent (Fig. 1). But due to differences in resolution and error between reanalysis data and numerical forecast data, there were certain deviations in local intensity and location. The results indicated that the gradient boosting model can effectively connect environmental field changes with the evolution process of lightning intensity, and had the ability for business prediction applications.

4 Conclusions and discussion

Based on historical lightning data in Ulanqab region and ERA5 environmental field data, a five-level classification prediction model for lightning intensity was constructed, which included 9 key environmental parameters. Among the four machine learning models evaluated, gradient boosting had the highest comprehensive score (0.173 7), the best overall predictive ability and stability, and its statistical advantages have been verified by DeLong test, confirming it as the final business prediction model. The simulation results of historical case showed that the model can better reflect the daily variation pattern of lightning activity increasing in the afternoon and decreasing in the evening, and effectively identify the evolution process of different intensity levels. The EC business prediction verification showed that the model can use future numerical forecast environmental fields to carry out lightning intensity prediction, and had the potential for business applications. Future work will focus on increase the sample size and integrate radar, satellite, and multi-source observation data to improve the ability to predict extreme lightning events.



Note: a_1, a_2, a_3, a_4 are the actual situation, b_1, b_2, b_3, b_4 are the ERA5 simulation, and c_1, c_2, c_3, c_4 are the EC prediction.

Fig. 1 Comparison of observed, ERA5-simulated, and EC-predicted lightning intensity distributions from 11:00 to 20:00 on 19 August 2024

References

[1] ZHAO Y, YAN JL, GUO YF, *et al.* Analysis of the relationship between the spatiotemporal distribution of land lightning and atmospheric environmental factors in China[J]. Chinese Journal of Atmospheric Sciences, 2025, 49(2): 313-324.

[2] ZHANG YB, CHENG L, LUO H, *et al.* Spatial and temporal distribution characteristics of cloud-to-ground lightning during the last decade in Shaanxi Province based on LLS[J]. Journal of Tropical Meteorology, 2017, 33(6): 884-895.

[3] CAO YC, ZHENG YG, SUN JS, *et al.* Spatiotemporal distributions and environmental characteristics of three types of regional severe convective weather processes associated with the Northeast China cold vortex[J]. Acta Meteorologica Sinica, 2024, 82(1): 22-36.

[4] ZHOU KH, ZHENG YG, YANG B, *et al.* Objective nowcasting of severe convective weather: Technological progress and outlook[J]. Acta Meteorologica Sinica, 2025, 83(3): 798-812.

[5] YANG J, CAO Z, DU Y, *et al.* Thunderstorm potential prediction based on random forests for the terminal area of the Guangzhou Baiyun International Airport[J]. Journal of Tropical Meteorology, 2022, 38(3): 387-396.

[6] LIU XW, HUANG WB, JIANG YS, *et al.* Study of the classified identification of the strong convective weathers based on the LightGBM algorithm[J]. Plateau Meteorology, 2021, 40(4): 909-918.

[7] ZHANG YF, FENG ZZ, LIU B. Lightning nowcasting early warning model based on convolutional neural network[J]. Meteorological Monthly, 2021, 47(3): 373-380.

[8] LI WJ, ZHAO F, LI MJ, *et al.* Forecasting and classification of severe convective weather based on numerical forecast and random forest algorithm[J]. Meteorological Monthly, 2018, 44(12): 1555-1564.

[9] LI WJ, LI MJ, MA H, *et al.* Severe convection prediction method based on XGBoost classified algorithm and numerical model ingredients[J]. Meteorological Monthly, 2024, 50(11): 1343-1358.

[10] CHEN JY, TANG DY, WU GS, *et al.* Lightning prediction capability improvement based on feature extraction and ensemble learning[J]. Journal of Tropical Meteorology, 2021, 37(3): 450-456.