ACCSCIENCE PUBLISHING

Asian Journal of Water, Environment and Pollution

ORIGINAL RESEARCH ARTICLE

Light gradient boosting machine-based early warning of dense fog in the Qiantang River Estuary with Shapley additive explanations interpretation

Nan Fang[1], Xiaoni Liang[2]*, Fengxue Qiao[3,4], Zhaoming Chen[3], Chuhan Lu[5], and Weicai Zheng[1]

[1]Early Warning Center in Zhejiang Province, Hangzhou, Zhejiang, China

[2]Meteorological Service Center in Zhejiang Province, Hangzhou, Zhejiang, China

[3]Department of Geography, School of Geographic Sciences, East China Normal University, Shanghai, China

[4]Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai, China

[5]Department of Atmosphere, School of Atmosphere and Remote Sensing, Wuxi University, Wuxi, Jiangsu, China

***Corresponding author:**
Xiaoni Liang
(lxn_zj@sina.com)

Citation: Fang N, Liang X, Qiao F, Chen Z, Lu C, Zheng W. Light gradient boosting machine-based early warning of dense fog in the Qiantang River Estuary with Shapley additive explanations interpretation. *Asian J Water Environ Pollut.* 2026;23(3):025500379. doi: 10.36922/AJWEP025500379

Received: December 8, 2025

Revised: February 7, 2026

Accepted: February 24, 2026

Published online: May 12, 2026

Abstract

Low-visibility events, particularly dense fog, pose significant risks to navigation and transportation safety in coastal estuarine regions, making accurate and timely early warning systems essential. This study develops a lightweight low-visibility warning model based on the light gradient boosting machine (LightGBM) algorithm, using hourly meteorological observation data from the Qiantang River Estuary region for the period 2021–2024. The forecast lead time was set to 3 h, and the model's performance was evaluated for predicting both low-visibility events (visibility ≤ 2,000 m) and fog events (visibility ≤ 500 m), with interpretability analysis conducted using Shapley additive explanations (SHAP). The results show that: (i) Validation against actual low-visibility events confirms that the model provided effective warnings across the study area, achieving an average accuracy of 98.6% for low-visibility events. (ii) The original LightGBM model requires parameter optimization to handle imbalanced classification, particularly for rare fog events. By adjusting class weights, the false-negative rate for dense fog was effectively reduced, improving recall from 44% to 70%. (iii) Global SHAP analysis revealed that relative humidity is the meteorological factor contributing most to dense fog warnings. Sample characteristics such as low wind speed, high humidity, and a small air–ground temperature difference consistently contribute to the model's prediction of dense fog events.

Keywords: Light gradient boosting machine; Visibility; Machine learning; SHapley Additive exPlanations analysis

1. Introduction

Atmospheric visibility is an important indicator of atmospheric transparency and air quality.[1] Low visibility can lead to flight delays and increased traffic accidents; consequently, low-visibility events have drawn significant attention in transportation, aviation, maritime navigation, and urban air-quality management.[2,3] Researchers have long been committed to studying the predictability of visibility to provide a scientific basis for traffic safety and urban air pollution control. Currently, operational visibility forecasts primarily rely on numerical weather prediction (NWP) models. However, constrained by inadequate parameterization schemes and limited grid resolution, NWP models often fail to accurately simulate small-scale, short-lived fog events.[4,5] This results in insufficient forecasting accuracy for localized low-visibility phenomena, such as radiation fog and advection fog, particularly over regions with complex terrain.[6]

In recent years, with the expansion of observational datasets and the decreasing cost of computational resources, machine learning methods have attracted widespread attention and achieved significant advances, becoming an important approach in statistics.[7] Compared with traditional NWP models, machine learning offers advantages such as smaller model size, lower computational demands, high portability, and reduced operational and maintenance costs, demonstrating considerable potential for operational meteorological services. Research indicates that since 2015, the use of artificial intelligence models—including machine learning and deep learning—has surged in meteorological applications.[8] For instance, Wang *et al.*[9] applied a risk neural network to daily visibility forecasting. Wang[10] developed a visibility prediction model for Shanghai using the extreme gradient boosting (XGBoost) algorithm. Zhang *et al.*[11] evaluated five mainstream machine learning algorithms for forecasting $PM_{2.5}$ mass concentration and found that XGBoost delivered the best overall performance. Fang *et al.*[12,13] constructed single-station visibility and wind speed prediction models using long short-term memory (LSTM) networks, revealing that the length of the input time window significantly influences predictive accuracy. Castillo-Botón *et al.*[14] explored various machine learning methods for both classification and regression tasks in fog prediction, reporting that gradient boosting performed best for classification, while random forests yielded optimal results for regression. Ding *et al.*[15] developed an LSTM-based model to forecast $PM_{2.5}$ concentration, providing a viable framework for time-series prediction in air quality modeling.

Among various machine learning algorithms, gradient boosting decision tree (GBDT) methods—particularly light gradient boosting machine (LightGBM)[16]—have demonstrated distinct advantages in meteorological applications and have achieved promising results in recent studies. For example, Zhang *et al.*[17] combined satellite observations with a LightGBM model to develop a remote sensing-based prediction scheme for subsurface ocean temperature and salinity, and further analyzed the contribution of different input features to model predictions using interpretability techniques. Wang and Tan[18] developed an objective dense fog forecasting model using ERA5 reanalysis data, providing valuable guidance for predicting potential fog areas in southern Henan province. Wang[19] developed multiple LightGBM-based post-processing models to correct numerical forecasts, significantly improving the accuracy of atmospheric visibility predictions.

However, existing studies on the application of machine learning models in meteorological forecasting have predominantly focused on optimizing predictive accuracy, with limited attention paid to the interpretability of the relationship between predicted events and meteorological input features. This lack of interpretability hinders a clear understanding of the model's decision-making mechanisms, thereby limiting assessments of its reliability and stability.

To elucidate the decision-making mechanisms of machine learning models from the perspectives of global and local interpretability and feature interactions, this study employed the SHapley Additive exPlanations (SHAP) algorithm to analyze model predictions. As an emerging model interpretability framework, SHAP has been theoretically extended and optimized by Lundberg *et al.*,[20] and is increasingly adopted to address the aforementioned challenges. For instance, Hou *et al.*[21] applied interpretable machine learning to quantify drivers of haze pollution, using SHAP analysis to decompose the contributions of "meteorology–emissions–chemistry" factors to $PM_{2.5}$ concentrations at the event level, thereby revealing the underlying physicochemical mechanisms of air pollution. Madhushani[22] employed SHAP-based interpretability to uncover the dynamic influences of key drivers—including year, precipitation, forest cover, and snow water equivalent—on runoff, providing a trustworthy basis for flood early-warning systems. Guo *et al.*[23] developed a rapid assessment model for rainfall-induced landslide hazards using categorical boosting and SHAP. SHAP identified 24-h rainfall, slope, and normalized difference vegetation index as the top factors and revealed a rainfall–slope interaction threshold, providing an interpretable basis for mountain hazard warnings. Covert *et al.*[24] proposed a "conditional SHAP" framework to

address the distortions in explanations from traditional marginal SHAP when handling correlated features, demonstrating reliable explanations for high-accuracy models on both simulated and real medical data. Xiahou and Xiao[25] analyzed warm-season heavy rainfall predictors in Jiangxi using SHAP, verifying that SHAP explanations are consistent with synoptic meteorology principles and operational experience. Dong[26] proposed a correlation–machine learning–SHAP multi-module coupled model for predicting atmospheric pollutant concentrations, which serves as a paradigm for interpretability research in meteorology and environmental science.

In summary, the primary objectives of this study are to develop a lightweight visibility prediction model based on LightGBM; apply SHAP-based interpretability analysis to investigate the physical interpretability of machine learning predictions; quantitatively identify meteorological features that exhibit high contribution to low-visibility forecasting; evaluate the accuracy and reliability of the model in predicting low-visibility events; and explore the integration mechanism between machine learning and atmospheric physical processes. This work aims to provide a novel modeling framework for visibility prediction under complex meteorological conditions.

Section 2 describes the study area, the observational data, the feature variables used for model construction, and the preprocessing methods. Section 3 introduces the LightGBM model architecture, the SHAP interpretability approach, and the evaluation metrics. Section 4 presents a systematic analysis of feature-variable variations at representative stations, model optimization for imbalanced dense fog samples, and validation of the model's low-visibility early warning capability. Section 5 discusses the limitations of this study and outlines directions for future work. Section 6 provides the main conclusions.

2. Methodology

2.1. Study area

The study area was the Qiantang River Estuary and its vicinity, a region at the heart of Zhejiang's economic region. Within this study area lies the Jia-Shao Bridge, a vital transportation link across the Qiantang River. This bridge is unique globally, being the only super-large bridge built in such a hydrogeological environment, which is also known for having one of the world's three largest tidal bores. Studying this area allowed us to examine low-visibility weather characteristics specific to its complex estuarine terrain and to provide scientific guidance for extreme low-visibility warnings on critical bridge sections. This work aims to improve risk response and disaster mitigation capabilities, ensuring bridge traffic safety and supporting regional economic development.

2.2. Observational data

Meteorological data for this study were collected from seven national basic meteorological stations around the Qiantang River Estuary: Xiaoshan district (Hangzhou), Haining city (Jiaxing), Haiyan county (Jiaxing), Keqiao district (Shaoxing), Shangyu district (Shaoxing), Yuyao city (Ningbo), and Cixi city (Ningbo) (see Figure 1). Historical observations of conventional meteorological elements were used, including air temperature, surface temperature, relative humidity, wind speed, visibility, and precipitation. The visibility sensor used is a HY-35P forward-scattering visibility meter. Other meteorological variables were measured by automatic instruments with a temporal resolution of 1 h.

2.3. Preliminary feature selection

Through quality assessment of historical meteorological observations and consideration of operational requirements, seven meteorological variables were selected as input features: air temperature (T), ground temperature (GT), relative humidity (RHU), station pressure (PRS), 10-m wind speed (WS), precipitation (PRE), and visibility (VIS). Given that the influence of meteorological factors often exhibits time lags and pronounced diurnal variations, lagged values of these variables at lead times of 1, 3, 6, 9, and 12 h prior to the forecast time were also included as predictors. This approach captures the temporal evolution of meteorological conditions while avoiding excessive data redundancy. In total, 45 input features were compiled for model development (Table 1).

To leverage the strength of LightGBM in handling classification tasks and align with operational needs for low-visibility warnings, the original regression problem of visibility forecasting was reformulated as a binary classification task—specifically, whether visibility falls below a predefined threshold. Multiple low-visibility thresholds were tested, and the model was trained to predict, based on current inputs, whether the visibility at the forecast time would meet or fall below the warning criterion. For consistent performance evaluation, all experiments used a fixed forecast lead time of 3 h; that is, input features at time t (Table 1) were used to predict whether low visibility would occur at t + 3 h.

To evaluate model performance, data from 2021 to 2023 were used for training, and 2024 data served as the independent test set. Considering that low-visibility events in the study region predominantly occur during autumn and winter, the training dataset focused on the period from October to March of the following year.

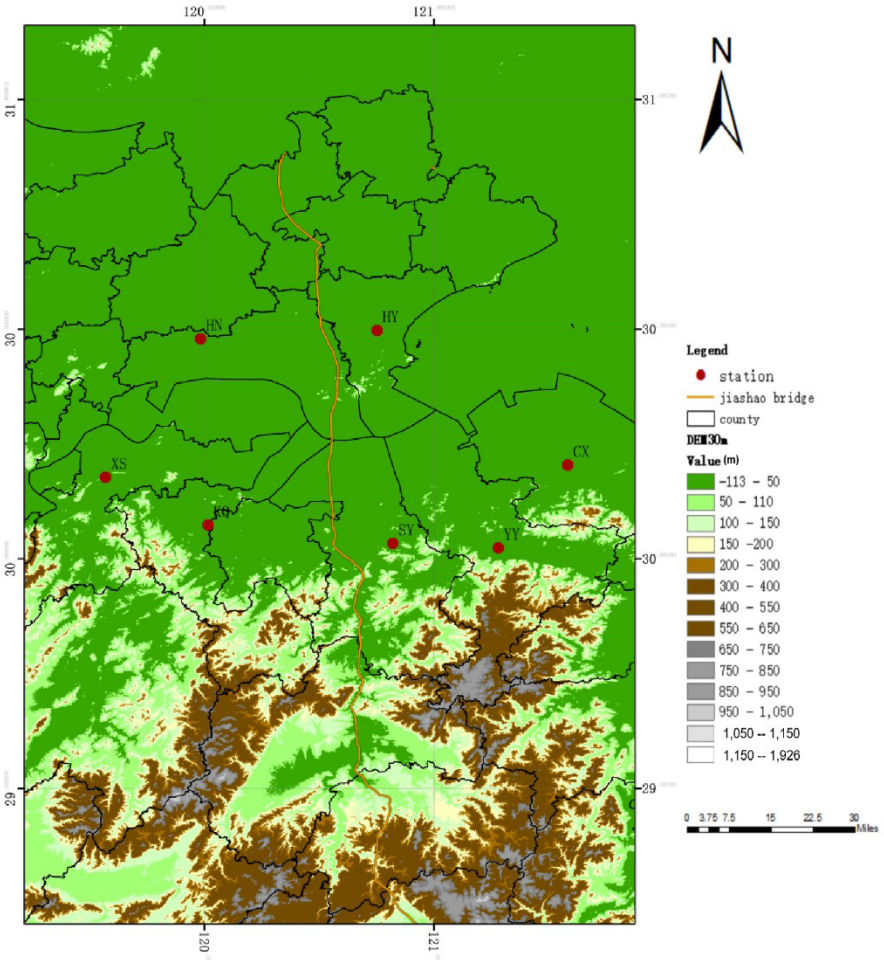

Figure 1. Map of the study area and locations of observation stations
Abbreviations: CX: Cixi; HN: Haining; HY: Haiyan; KQ: Keqiao; SY: Shangyu; XS: Xiaoshan; YY: Yuyao.

Table 1. Classification of feature variables

Temperature (°C; 18 variables)	Relative humidity (%; 6 variables)	Pressure (hPa) – 3 variables	Wind speed (m/s; 6 variables)	Precipitation (mm; 6 variables)	Visibility (m; 6 variables)
T: Current air temperature T_n: Air temperature n hours earlier GT: Current surface temperature GT_n: Ground temperature n hours earlier DTDG: Current air–ground temperature difference (temperature inversion intensity) DTDG_n: Air–ground temperature difference n hours earlier	RHU: Current relative humidity RHU_n: Relative humidity n hours earlier	PRS: Station pressure PRS_3: Three-hour pressure change PRS_24: 24-h pressure change	WS: Current wind speed WS_n: Wind speed n hours earlier	PRE_1: Precipitation in the past hour PRE_n: Accumulated precipitation over the past n hours	VIS: Current visibility VIS_n: Visibility n hours earlier

Note: n = 1, 3, 6, 9, 12 (for variables with _n suffix).

2.4. Data preprocessing

Using the aforementioned approach, a training dataset for an hourly dense fog warning model covering 2021–2024 was constructed. Since the meteorological observational data selected for this study originated from national basic weather stations, the quality is high, and there were no prolonged missing intervals. Any isolated outliers or missing data points were processed with linear interpolation to ensure the completeness of the dataset.

Due to the different units and large numerical ranges of the meteorological variables, the proper functioning of the loss function can be hindered. To ensure that each variable's impact on visibility was treated equally, we rescaled predictors via min–max normalization, scaling each variable to the range [0, 1]. The normalization equation is:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x' is the normalized value, x is the original value, while x_{min} and x_{max} are the minimum and maximum values of that variable over the sample period, respectively.

For model validation, the dataset was partitioned using stratified five-fold cross-validation. This method enhances standard five-fold cross-validation, making it ideal for classification tasks with imbalanced class distributions. It ensures each fold maintains the same class proportion as the whole dataset, preventing any fold from missing rare events and yielding fairer, more stable evaluations.

2.5. Description of the light gradient boosting machine model

The light gradient boosting machine is an open-source gradient boosting tree algorithm developed by Microsoft that provides an efficient distributed decision-tree framework. Therefore, LightGBM delivers a faster, more accurate, and more scalable machine learning algorithm, demonstrating excellent performance across various datasets. For this reason, this study employed this algorithm to construct a visibility prediction model. After evaluation and comparison, the final LightGBM model parameters adopted are as shown in Table 2.

2.6. SHapley Additive exPlanations interpretability approach

The SHAP is a post-hoc additive framework that assigns each feature a Shapley value—the feature's contribution to the prediction relative to baseline, solving the problem of fairly distributing rewards among cooperative players. In machine learning, we treat input features as "players" and the model prediction as the "payout." SHAP values quantify each feature's contribution to the model's output, linking the predicted outcome to the input features in a principled, quantitative way.

2.7. Model evaluation metrics

To objectively quantify the model's performance, we selected three metrics as evaluation criteria well-suited for classification problems: accuracy, precision, and recall. With stratified five-fold cross-validation during training, we reported the average of the five folds for each metric

Table 2. Main parameters of the light gradient boosting machine model

Parameter name	Parameter description	Parameter value
Objective	Specifies the learning task as binary classification (labels are 0 or 1)	Binary
Boosting_type	Uses the standard gradient boosting decision tree (GBDT) algorithm	GBDT
Num_leaves	Maximum number of leaves per tree	31
Class_weight	Assigns different weights to classes	{0: 1, 1: 10}
N_estimators	Builds 500 decision trees	500
Learning_rate	Learning rate (shrinkage) that scales the contribution of each tree	0.05
Min_child_samples	Minimum number of samples required in a leaf node	20

as the final performance. In classification tasks, there are four possible outcomes for each prediction. In the context of low-visibility warnings, they are defined as:

- True positive (TP): Correctly predicted a low-visibility event.
- False positive (FP): Incorrectly predicted a low-visibility event (a false alarm).
- False negative (FN): Failed to predict a low-visibility event (a missed event).
- True negative (TN): Correctly predicted clear skies.

Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

It is the proportion of all correctly predicted instances (TP + TN) out of the total instances.

Dense fog events are inherently rare in practice, leading to an imbalanced training dataset. This imbalance can make overall accuracy obscure the model's true performance on fog events. Given that missed dense fog events (FN) were of particular concern due to their serious consequences, as such events frequently lead to traffic accidents, flight delays, and adverse public health impacts, thereby causing substantial socioeconomic losses, it is critical to pay special attention to both precision and recall for the dense fog events:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

3. Results and discussion

3.1. Influencing factors for the low-visibility warming model

To characterize the meteorological conditions during low-visibility events, this study used hourly data from the Shangyu station over 2021–2024 as an example, analyzing the monthly mean duration of low-visibility event occurrence at each hour and the hourly averages of relative humidity, wind speed, 3-h pressure change, and air–ground temperature difference (DTDG) for each month.

Figure 2 shows that dense fog predominantly occurred in the late night and early morning hours (around 23:00–24:00 and 00:00–06:00) during winter and spring, while being rare in the afternoon (12:00–18:00) and during summer and early autumn (July–October). Among the meteorological conditions, relative humidity exhibited a clear diurnal cycle, being generally low during the day and high at night. Wind speed was typically higher in the afternoon (12:00–18:00), with summer afternoons having the strongest winds, whereas the lowest wind speeds occurred in the early morning (00:00–06:00). The 3-h pressure change showed a distinct "dual peaks and dual valleys" pattern, with positve pressure changes in the morning and early night, negaive changes in the afternoon, and change little (weakly negative) in the late night to early morning. The DTDG, a key indicator of near-surface atmospheric inversion structure, also showed a pronounced diurnal variation: during the day (08:00–16:00), air temperature was lower than ground temperature, and vice versa at night.

In summary, low-visibility events consistently occur under conditions of high relative humidity and low wind speed, as observed in previous studies. Additionally, these events are more likely to occur when the air temperature exceeds the ground temperature (DTDG > 0 °C), consistent with a shallow near-surface inversion. The other six stations showed similar patterns (data not shown).

3.2. Model optimization for dense fog events under imbalanced conditions

Low-visibility events were rare: samples with visibility ≤ 2,000 m accounted for only 11.2% of the dataset, and those with visibility ≤ 500 m (dense fog) constituted merely 1.7%. The baseline LightGBM model tended to be overly conservative in predicting such low-probability events. Although it achieved a high overall accuracy (ACC) of 97.5%, its recall for dense fog events was only 44%, indicating that more than half of all dense fog cases were missed.

Previous studies have proposed various strategies to mitigate the impact of class imbalance, which can be broadly categorized into three types: undersampling, oversampling, and hybrid approaches.[27] In this study, we first leveraged domain knowledge to partially alleviate imbalance: as shown in the previous section, low-visibility events in the study region predominantly occurred during nighttime to early morning hours (23:00–08:00) from November to April. By restricting model training to data from this high-risk period, we effectively reduced the proportion of non-fog samples and improved class balance.

Another effective approach is to assign class weights during model training. In LightGBM, each sample's loss is scaled by the weight of its corresponding class during gradient computation. Consequently, misclassifying a sample from a high-weight class incurs a larger penalty in the total loss function. Initially, we applied LightGBM's

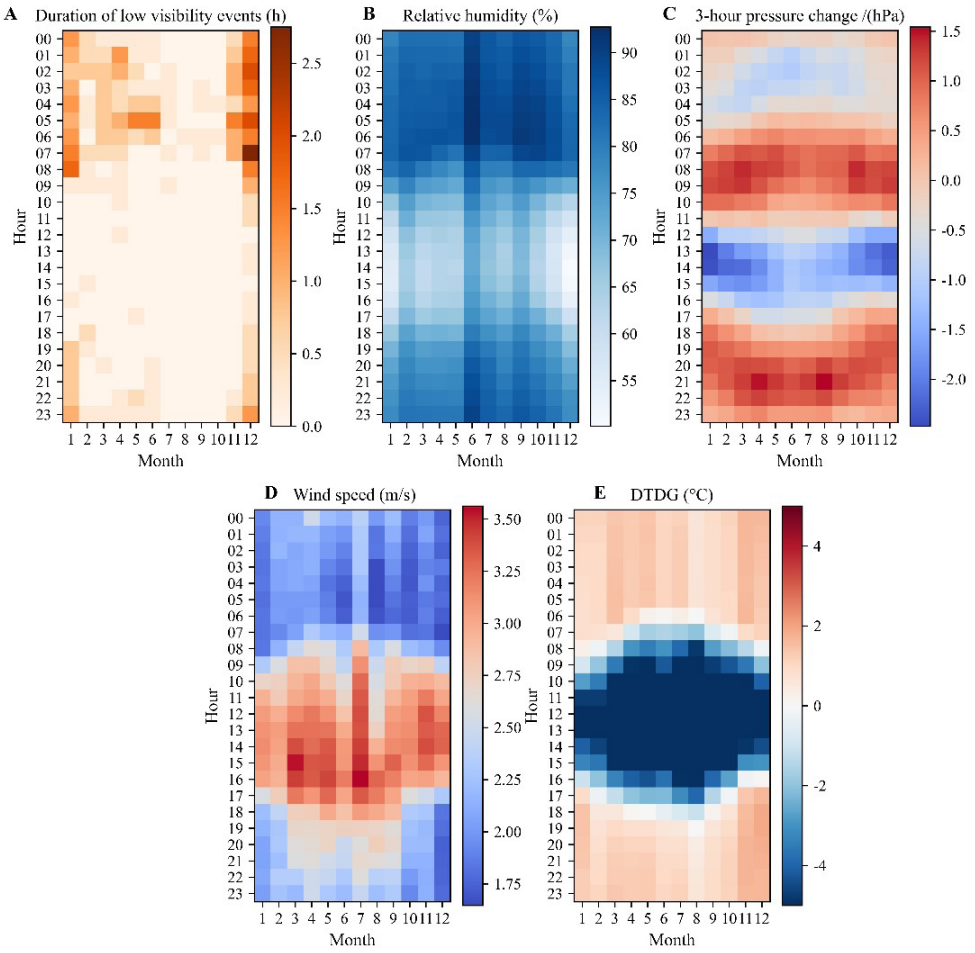

Figure 2. Hourly variations of meteorological variables for each month during 2021–2024 at Shangyu station. (A) Duration of low visibility events. (B) Relative humidity. (C) 3-h passage pressure. (D) Wind speed. (E) Air–ground temperature difference.

built-in scale_pos_weight parameter to automatically balance classes; this increased dense fog recall to 80%, but at the cost of precision, which dropped from 79% to 45%—indicating a significant rise in false alarms.

This trade-off suggests that the default balancing strategy is suboptimal for operational forecasting, where both missed detections and false alarms carry substantial costs. To address this, we conducted a systematic grid search over class weight ratios. We found that setting the weight ratio between non-dense fog and dense fog samples to 1:10 yielded the best overall performance: dense fog recall improved from 44% to 70%, while maintaining acceptable precision. Similarly, for models targeting other low-visibility thresholds, optimal class weight ratios were individually tuned to maximize predictive skill under operational constraints (Figure 3).

3.3. Case-based validation of the low-visibility warning model

To comprehensively evaluate the performance of the low-visibility warning model, we conducted case studies using low-visible events from the validation dataset for each of the seven observation stations. The selected low-visibility episodes included: Xiaoshan (XS) on October 25–26; Haining (HN) on November 3–4; Haiyan (HY) on December 10–11; Keqiao (KQ) on November 15–16; Shangyu (SY) on December 10–11; Cixi (CX) on November 12–13; and Yuyao (YY) on October 26–27. The warning threshold was set to VIS ≤ 2,000 m.

Figure 4 shows the model's low-visibility warning outputs versus observed visibility at all seven sites. The blue curve depicts the observed hourly visibility at each station, while the red vertical bars indicate times when the

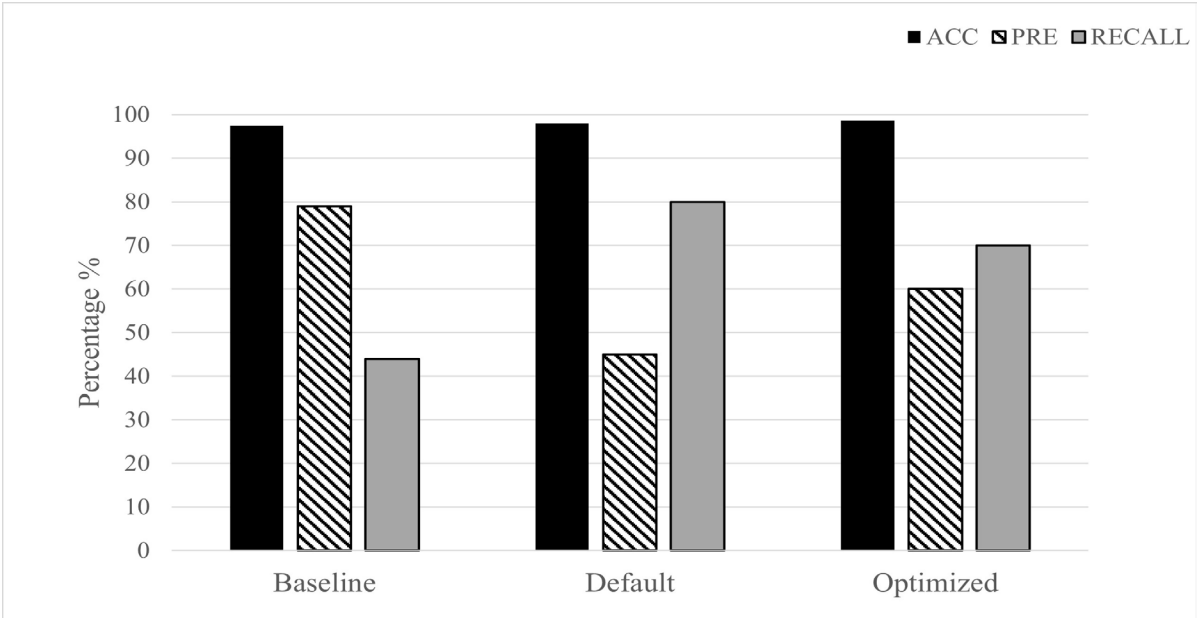

Figure 3. Evaluation results of the dense fog warning model (optimized vs. baseline)
Abbreviations: ACC: Accuracy; PRE: Precipitation.

model issued a low-visibility warning. Overall, the model performed well in all seven regions and even captured the timing of intermittent visibility improvements. The recall for low-visibility events exceeded 90% across all seven stations, with HY, KQ, SY, and CX achieving perfect recall (100%), indicating no missed low-visibility events at these sites. However, the precision at HY, KQ, and SY was relatively low, suggesting a higher false-alarm rate. In summary, the main shortcoming is that the model is not sufficiently sensitive to threshold precision: when visibility is marginal (2,000–4,000 m, just above the warning criterion), it still issues a low-visibility warning.

3.4. Interpretability analysis based on Shapley additive explanations values

Machine learning-based prediction models offer advantages such as convenient data acquisition and efficient computation, but they generally lack physical interpretability. Therefore, it is essential to evaluate the interpretability of machine learning models for dense fog warning.

3.4.1. Shapley additive explanations analysis of the dense fog warning model

Figure 5 presents a SHAP summary plot for dense fog events (VIS ≤ 500 m), with features ranked by their mean absolute SHAP value from a stratified five-fold cross-validation. On the plot, the x-axis shows SHAP values: a positive value indicates that the feature increases the probability of a dense fog event, while a negative value indicates it decreases that probability. The y-axis lists the features in descending order of importance. Each point represents one sample, and the point's color indicates the actual value of that feature for that sample (e.g., red for higher values, blue for lower). Relative humidity was the most influential factor driving the model's dense fog predictions: samples with high relative humidity are strongly pushed toward the dense fog class. The second most influential factor was DTDG. A positive DTDG (meaning the air temperature is slightly higher than the ground temperature) contributes positively to the model's dense fog prediction. However, it is worth noting that the positive contribution of the DTDG did not increase monotonically with its magnitude. Instead, the strongest positive contribution occurred when the air temperature was slightly higher than the ground temperature. This relationship was quantitatively analyzed in detail in the following section. Additionally, features reflecting low temperature, low pressure, and low wind speed also contributed positively to the model's prediction of dense fog events. A comparison of the contribution rankings across input features revealed that meteorological variables from recent time steps were more important and ranked higher, whereas those from earlier time steps generally ranked lower. This indicates that, for dense fog prediction, meteorological conditions in the immediate past play a decisive role.

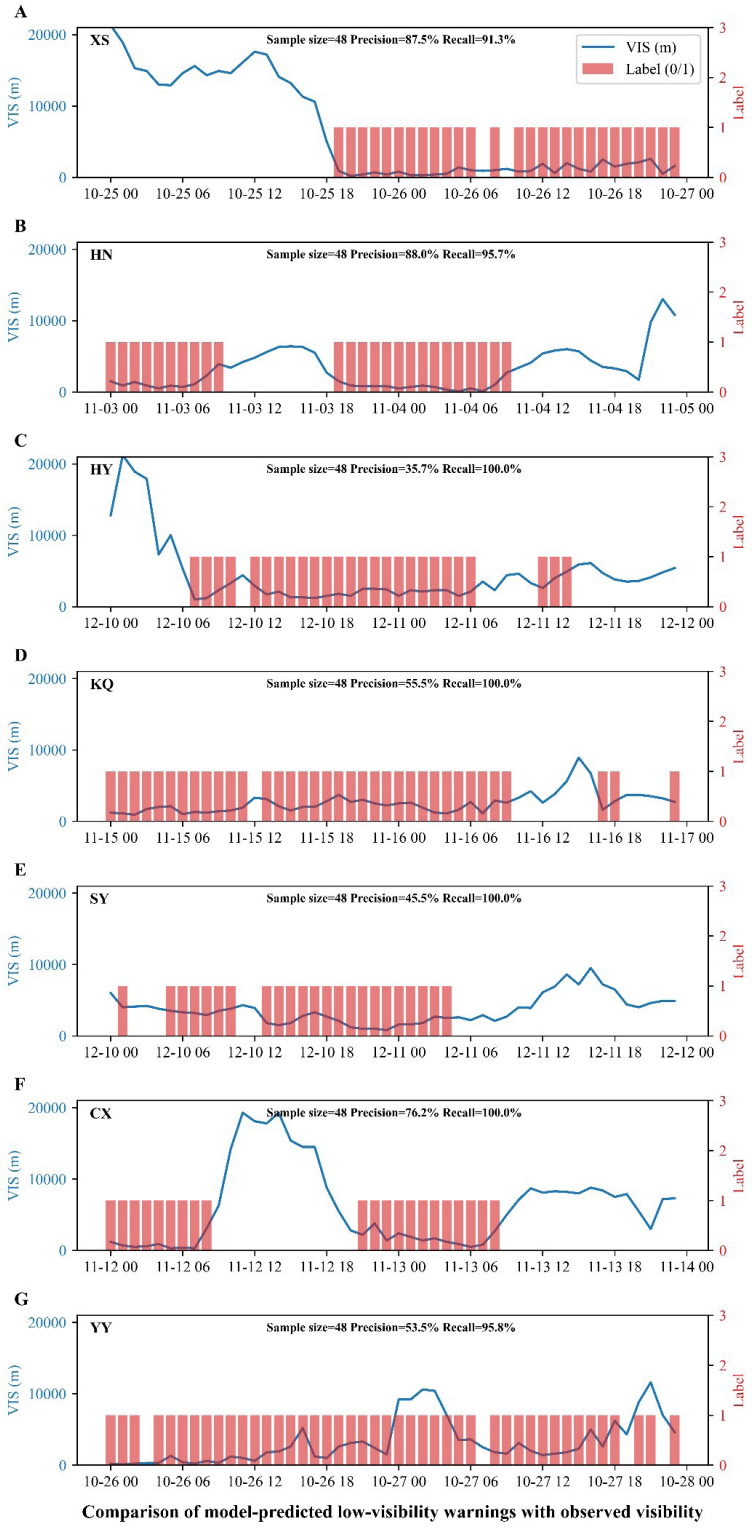

Figure 4. Comparison of model-predicted low-visibility warnings with observed visibility at the seven stations: (A) Xiaoshan (XS), (B) Haining (HN), (C) Haiyan (HY), (D) Keqiao (KQ), (E) Shangyu (SY), (F) Cixi (CX), and (G) Yuyao (YY).
Abbreviation: VIS: Visibility.

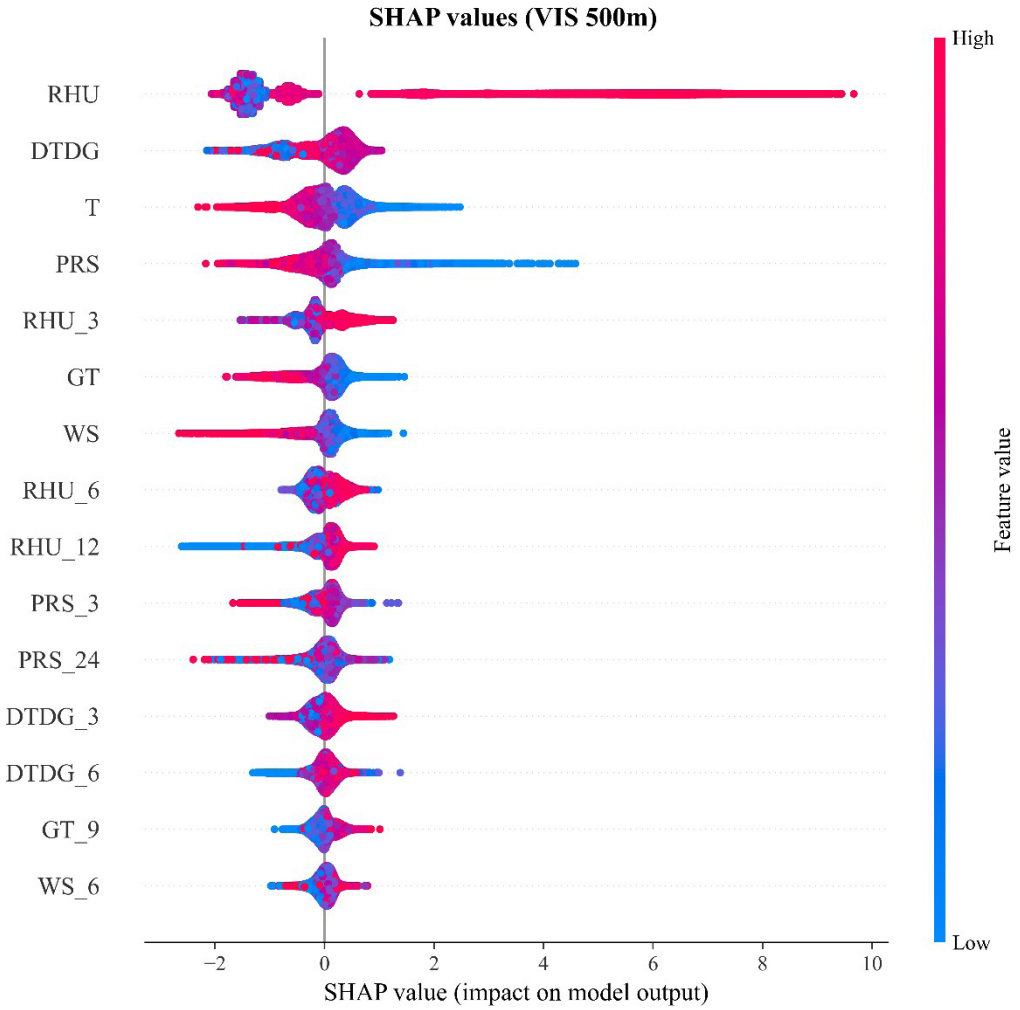

Figure 5. Distribution of feature contributions for the dense fog warning model (Shapley additive explanations [SHAP] summary plot)
Abbreviations: DTDG: Air–ground temperature difference; GT: Ground temperature; PRS: Station pressure; RHU: Relative humidity; T: Air temperature; WS: 10-m wind speed.

3.4.2. Interaction effects of meteorological factors on dense fog events

Analyzing interactions among the top contributing factors allows for a more objective and quantitative understanding of their contribution mechanisms to the dense fog warning model, as shown in Figure 6. For instance, when air and ground temperatures were nearly equal (DTDG ≤ 2 °C) and relative humidity exceeded 90%, the model was more likely to predict a dense fog event. When the air temperature was below 10 °C, and the ground temperature was below 5 °C, these conditions also favored dense fog predictions. Likewise, samples with wind speed < 2 m/s, relative humidity ≥ 90%, and a very small DTDG (≤ 2 °C) collectively contributed positively to predicting dense fog events.

Overall, samples characterized by wind speed < 2 m/s, relative humidity ≥ 90%, air temperature < 10 °C, ground temperature < 5 °C, and a small DTDG (≤2 °C) consistently exhibited positive contributions to the model's prediction of dense fog events.

These key features align closely with the known physical mechanisms of dense fog formation in this region. Radiation fog—the most common fog type during winter—typically forms under clear skies, light winds, and high humidity from nighttime to early morning. The Qiantang River Estuary is extensively covered by water bodies, which supply abundant moisture through evaporation in winter. Intense longwave radiative cooling of the surface rapidly cools the near-surface air, leading to condensation and fog formation.

3.4.3. Meteorological factor contributions under varying thresholds

A different threshold for defining a dense fog event can significantly alter the importance ranking of meteorological factor contributions. Figure 7 compares the contribution distributions of features when defining fog events at visibility thresholds of 500 m, 1,000 m, 1,500 m, and 2,000 m. Under the most extreme fog conditions (VIS ≤ 500 m), features from the immediate past (current conditions) contributed the most to predictions. As the threshold rose (less extreme events), features reflecting historical conditions gradually increased. These include features such as relative humidity over the past 3 and 6 h, the 24-h pressure change, and wind speed from the past 9 and 12 h. This suggests that for predicting the occurrence of general low-visibility events, the model has stronger lead-time predictability, and the overall synoptic weather pattern has a larger influence, whereas extreme dense fog events depend more on immediate meteorological conditions.

4. Limitations and future work

Considering operational practicality, this study constructed the dataset using only historical observational data. This approach significantly simplifies data preparation and reduces storage requirements, making it well-suited for rapidly deploying lightweight, single-station low-visibility warning models in operational settings. However, the potential impacts of data latency or missing observations were not thoroughly addressed in this work. In future research, we plan to incorporate numerical model reanalysis data, operational forecast outputs, and other complementary data sources; develop robust strategies to handle missing inputs; and enrich the feature space to further investigate the contribution of diverse predictors.

Moreover, our experiments reveal that the current model exhibits limited sensitivity in forecasting extremely dense fog events. To address this, we will explore a broader range of hyperparameter configurations and alternative machine learning architectures to comprehensively evaluate the

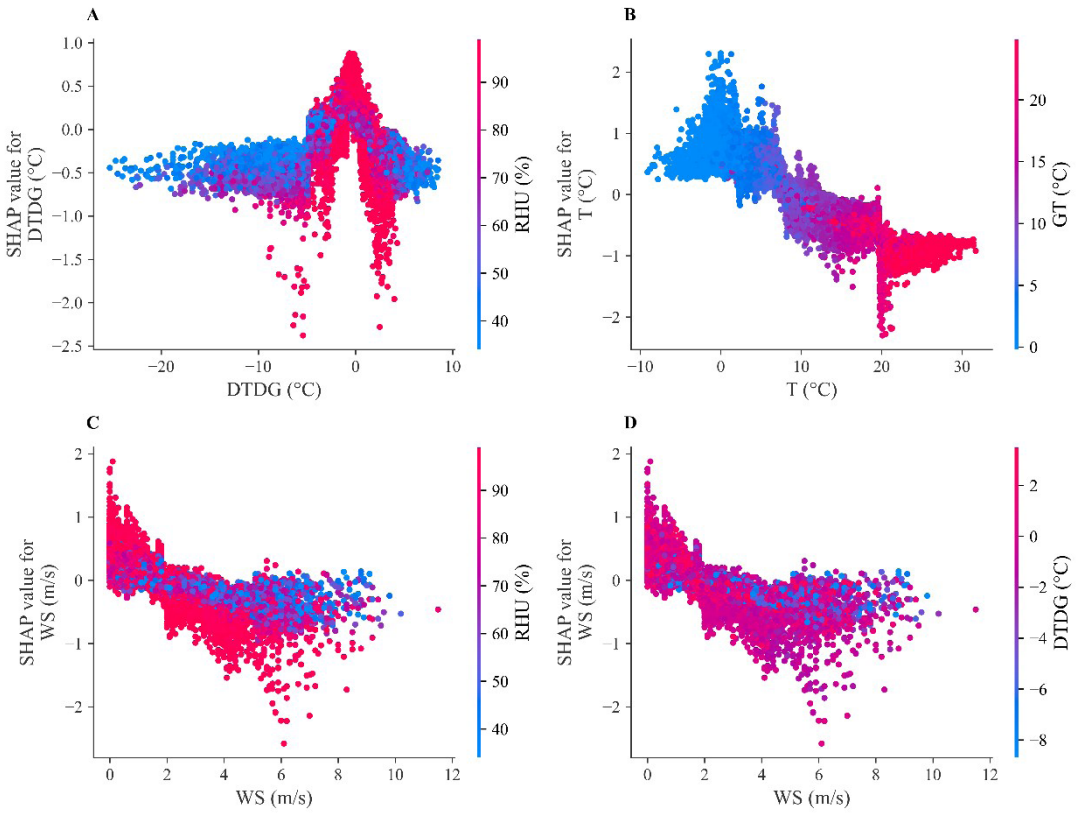

Figure 6. Effects of interactions among multiple meteorological factors on dense fog event prediction
Abbreviations: DTDG: Air–ground temperature difference; GT: Ground temperature; RHU: Relative humidity; SHAP: Shapley additive explanations; T: Air temperature; WS: 10-m wind speed.

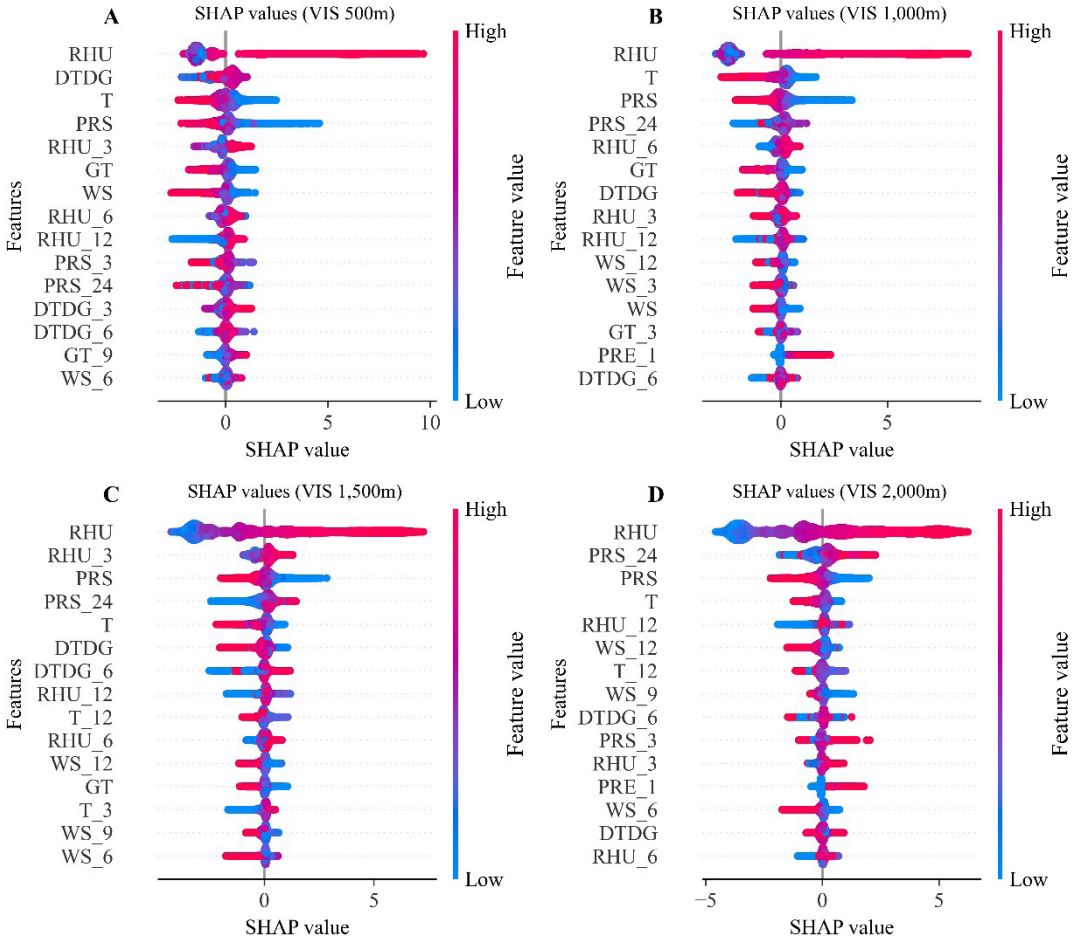

Figure 7. Feature contribution comparison under different dense fog threshold conditions
Abbreviations: DTDG: Air–ground temperature difference; GT: Ground temperature; PRS: Station pressure; RHU: Relative humidity; SHAP: Shapley additive explanations; T: Air temperature; VIS: Visibility; WS: 10-m wind speed.

potential of machine learning-based approaches for early warning of extreme low-visibility conditions.

5. Conclusion

This study used hourly meteorological observations from seven sites in the Qiantang River Estuary area from 2021 to 2024 to construct a lightweight low-visibility warning model based on LightGBM. By investigating interactions among meteorological variables, the feature combinations for the model's input data were optimized to enhance its predictive performance in dense fog warning. The SHAP interpretability method was specifically employed for dense fog events with varying visibility thresholds to explore key influencing factors, validate the rationality of low-visibility predictions, and better quantitatively understand the contribution mechanisms of different factors to the dense fog warning model. The main conclusions are as follows:

(i) The LightGBM model shows skill for low-visibility warnings at a 3-h lead time. However, for imbalanced classification problems, parameter tuning is required. Adjusting class weights can effectively reduce the model's miss rate for dense fog events. Comparison with actual low-visibility occurrences shows that the weight-optimized model provides effective warnings across the study area. The overall accuracy of low-visibility warnings was 98.6%; for dense fog events (VIS ≤ 500 m), the precision was 60%, and the recall was 70%.

(ii) SHAP analysis indicates that relative humidity is the most influential meteorological factor for dense fog predictions. Samples characterized by low wind speed (< 2 m/s), high humidity (RH ≥ 90%), and a near-zero DTDG (≤ 2 °C) contribute positively to the model's dense fog predictions.

(iii) Analysis of feature importance under different warning thresholds shows that the model has stronger predictive capability for general low-visibility events (VIS ≤ 2,000 m) than for extreme dense fog (VIS ≤ 500 m), with a greater contribution from the overall synoptic situation. By contrast, extremely dense fog events are more closely determined by immediate local meteorological factors.

Acknowledgments

None.

Funding

This research was funded by Zhejiang Provincial Natural Science Foundation of China (Grant No. LZJMY25D050007).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Nan Fang, Xiaoni Liang
Formal analysis: Nan Fang
Investigation: Nan Fang, Xiaoni Liang, Chuhan Lu
Methodology: Xiaoni Liang, Weicai Zheng
Writing–original draft: Nan Fang, Fengxue Qiao
Writing–review & editing: Fengxue Qiao, Zhaoming Chen

Availability of data

The data are available from the corresponding author upon reasonable request.

References

1. Fu G, Li X, Wei N. Review on the atmospheric visibility research. *Period Ocean Univ China*. 2009;39(5):855-862. In Chinese.

2. Guijo-Rubio D, Gutiérrez PA, Casanova-Mateo C, Sanz-Justo J, Salcedo-Sanz S, Hervás-Martínez C. Prediction of low-visibility events due to fog using ordinal classification. *Atmos Res*. 2018;214:64-73.

   doi: 10.1016/j.atmosres.2018.07.017

3. Qian WH, Leung JCH, Chen YL, Huang SY. Applying anomaly-based weather analysis to the prediction of low visibility associated with the coastal fog at Ningbo-Zhoushan Port in East China. *Adv Atmos Sci*. 2019;36:1060-1077.

   doi: 10.1007/s00376-019-8252-5

4. Zhou G, Xu J, Xie Y, *et al*. Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. *Atmos Environ*. 2017;153:94-108.

   doi: 10.1016/j.atmosenv.2017.01.020

5. Xia F, Li CY. Fog weather forecast experiments over Shandong province based on three visibility schemes. *J Meteorol Environ*. 2018;34(3):48-57. In Chinese.

   doi: 10.3969/j.issn.1673-503X.2018.03.006

6. Cheng FY, Feng CY, Yang ZM, *et al*. Evaluation of real-time PM2.5 forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM2.5 forecasts in Taiwan. *Atmos Environ*. 2021;244:117909.

   doi: 10.1016/j.atmosenv.2020.117909

7. Kurt A, Oktay AB. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Syst Appl*. 2010;37(12):7986-7992.

   doi: 10.1016/j.eswa.2010.05.093

8. Salcedo-Sanz S, Guijo-Rubio D, Pérez-Aracil J, Peláez-Rodríguez C, Gómez-Orellana AM, Gutiérrez-Peña PA. Artificial Intelligence-Based Methods and Algorithms in Fog and Atmospheric Low-Visibility Forecasting. *Atmosphere*. 2025;16(9):1073.

   doi: 10.3390/atmos16091073

9. Wang K, Zhao H, Liu AX, Han B, Bai Z. Development and validation of visibility forecast technique based on the risk neural network. *China Environ Sci*. 2009;29(10):1029-1033. In Chinese.

   doi: 10.3321/j.issn:1000-6923.2009.10.005

10. Wang Y. *Research on Shanghai Visibility Prediction Model Based on Multi-source Data and XGBoost Algorithm*. Dissertation. East China Normal University; 2019.

11. Zhang XT, Liu H, Liang M, Ju F, Gao XX. Prediction and analysis of PM2.5 mass concentration in Fenwei Plain based on different algorithms of machine learning. *J Shaanxi Meteorol*. 2023;(3):8-16. In Chinese.

    doi: 10.3969/j.issn.1006-4354.2023.03.002

12. Fang N, Jiang SJ, Yan XM, Ruan SJ, Ma XY. Research on Ultra Short-Term Fast Rolling Prediction Technology of Wind Speed Based on LSTM Neural Network. *Meteorol Sci Technol*. 2022;50(6):842-850. In Chinese.

    doi: 10.19517/j.1671-6345.20220064

13. Fang N, Xie GQ, Ruan XJ, Ren CP, Jiang SJ, Zhang WW. Application of long short-term memory neural network (LSTM) model in low visibility forecast. *J Meteorol Environ*. 2022;38(5):34-41. In Chinese.

    doi: 10.3969/j.issn.1673-503X.2022.05.004

14. Castillo-Botón C, Casillas-Pérez D, Casanova-Mateo C, *et*

al. Machine learning regression and classification methods for fog events prediction. *Atmos Res*. 2022;272:106157.

doi: 10.1016/j.atmosres.2022.106157

15. Ding YX, Li Z, Zhang CD, Ma J, Cheng JCP, Wan Z. Prediction of ambient PM2.5 concentrations using a correlation filtered spatial-temporal long short-term memory model. *Appl Sci*. 2019;10(1):14.

doi: 10.3390/app10010014

16. Ke G, Meng Q, Finley T, *et al*. LightGBM: A highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*; *December 4-9, 2017; Long Beach, CA, USA*. Neural Information Processing Systems Foundation; 2017:3146-3154. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf [Last accessed on December 8, 2025].

17. Zhang TY, Su H, Yang X, Yan XH. Remote sensing prediction of global subsurface thermohaline and the impact of longitude and latitude based on LightGBM. *Nat Remote Sens Bull*. 2020;24(10):1255-1269.

doi: 10.11834/jrs.20200007

18. Wang LL, Tan JH. Objective forecast research of dense fog in southern Henan based on LightGBM. *Arid Meteorol*. 2024;42(5):702-709. In Chinese.

doi: 10.11755/j.issn.1006-7639-2024-05-0702

19. Wang ZY. *The Correction of Atmospheric Visibility Prediction in Shanghai Based on LightGBM Framework*. Dissertation. East China Normal University; 2019.

20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*; *December 4-9, 2017; Long Beach, CA, USA*. Neural Information Processing Systems Foundation; 2017:4768-4777. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [Last accessed on December 8, 2025].

21. Hou L, Dai Q, Song C, *et al*. Revealing drivers of haze pollution by explainable machine learning. *Environ Sci Technol Lett*. 2022;9(2):112-119.

doi: 10.1021/acs.estlett.1c00865

22. Madhushani C, Dananjaya K, Ekanayake IU, Meddage DPP, Kantamaneni K, Rathnayake U. Modeling streamflow in non-gauged watersheds with sparse data considering physiographic, dynamic climate, and anthropogenic factors using explainable soft computing techniques. *J Hydrol*. 2024;631:130846.

doi: 10.1016/j.jhydrol.2024.130846

23. Guo ZZ, He J, Huang D, Zhou YQ, Zhu YH. Fast assessment model for rainfall-induced shallow landslide hazard and application. *Chin J Rock Mech Eng*. 2023;42(5):1188-1201. In Chinese.

doi: 10.13722/j.cnki.jrme.2022.0605

24. Covert I, Lundberg S, Lee SI. Explaining by removing: A unified framework for model explanation. *J Mach Learn Res*. 2021;22(33):1-90. Available from: https://www.jmlr.org/papers/volume22/20-1316/20-1316.pdf [Last accessed on].

25. Xiahou J, Xiao A. Importance analysis on warm season rainstorm forecast factors in Jiangxi province based on machine learning model of Shapely values. *Meteorol Disaster Reduct Res*. 2024;47(1):12-23. In Chinese.

doi: 10.12013/qxyjzyj2024-002

26. Dong JQ. *Prediction of atmospheric pollutant concentrations and driver factor mining based on interpretable machine learning*. Dissertation. North China Electric Power University; 2023.

27. Rathnayake N, Jayasinghe J, Semasinghe R, Rathnayake U. Predicting short-term wind power generation at Musalpetti Wind Farm: Model development and analysis. *Comput Model Eng Sci*. 2025;143(2):2287-2305.

doi: 10.32604/cmes.2025.064464