




## Full length article

# Estimating atmospheric visibility and PM from Doppler wind lidar with traditional and machine learning models

Wenhan Xu<sup>a, </sup>, Tianwen Wei<sup>a,\*, </sup>, Mengya Wang<sup>a</sup>, Zhekai Li<sup>a</sup>, Zhuoqun Li<sup>a</sup>, Zhen Zhang<sup>a</sup>, Haiyun Xia<sup>a, </sup>

<sup>a</sup> State Key Laboratory of Climate System Prediction and Risk Management (CPRM), School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing 210044, China

<sup>b</sup> School of Earth and Space Science, University of Science and Technology of China, Hefei 230026, China

## ARTICLE INFO

## Keywords:

Doppler wind lidar  
Visibility estimation  
Particulate matter  
Backscatter coefficient  
Machine learning

## ABSTRACT

Doppler wind lidar is widely used for wind profiling, while its potential for aerosol-based visibility and particulate matter (PM) estimation remains underexplored. This study assesses the capability of Doppler lidar backscatter measurements, combined with meteorological variables, for estimating visibility, PM<sub>2.5</sub>, and PM<sub>10</sub> concentrations. Traditional regression models were benchmarked against machine learning (ML) approaches, including tree-based, vector-based and neural-based methods. A logarithmic-linear empirical function was applied for visibility estimation, while PM estimation used a multivariate regression scheme accounting for the dependence of relative humidity due to the hygroscopic growth effects. Among ML methods, Light Gradient Boosting (LGB) consistently achieved the best performance across all tasks, yielding test-set results for visibility with RMSE of 3.38 km and R<sup>2</sup> of 0.85; for PM<sub>2.5</sub> with RMSE of 7.02 µg/m<sup>3</sup> and R<sup>2</sup> of 0.83; and for PM<sub>10</sub> with RMSE of 15.32 µg/m<sup>3</sup> and R<sup>2</sup> of 0.81. Feature importance analysis revealed that the backscatter coefficient was the dominant predictor for visibility (39.42 %), while the backscatter coefficient also proved most crucial for PM<sub>2.5</sub> (23.25 %), and relative humidity was critical for PM<sub>10</sub> estimation (26.19 %). These results demonstrate the capability of CDL measurements to estimate visibility and particulate matter concentrations. With their increasing deployment, this approach significantly extends the application of Doppler lidars to comprehensive air quality monitoring.

## 1. Introduction

Atmospheric aerosols form a complex multiphase system consisting of suspended solid and liquid particles. They play a crucial role in radiative transfer and atmospheric chemistry. Among these suspended particulates, particulate matter (PM) with aerodynamic diameter below 10 µm (PM<sub>10</sub>) and 2.5 µm (PM<sub>2.5</sub>) has emerged as a critical environmental metric due to its size-dependent atmospheric residence time and penetration capacity into human respiratory systems [1]. Beyond direct health impacts [2], PM exerts optical effects through Mie scattering, becoming the dominant light extinction agent in polluted atmospheres [3,4]. This scattering-induced visibility degradation, quantified by meteorological optical range (MOR) [5], poses significant risks to aviation safety and transportation efficiency [6,7]. The PM-visibility relationship is governed by aerosol microphysical properties: particle

size distribution dictates scattering cross-sections, chemical composition influences hygroscopic growth under varying humidity conditions, and mixing states modulate light absorption [8,9,10]. These interrelated properties highlight PM's dual role as both an air pollutant and a key modulator of atmospheric optics, underscoring the need for integrated monitoring of PM concentrations and visibility variations.

While precise, conventional PM monitoring via β-ray absorption and gravimetric methods is restricted to fixed stations, a major limitation given PM's dynamic vertical transport and horizontal advection [11,12,13]. Similarly, transmissive and scattering visibility meters provide point measurements that are inadequate for resolving the three-dimensional extinction field [14,15,16]. These spatial and temporal limitations drive the need for remote sensing solutions, including passive and active methods. Satellite monitoring, as a passive remote sensing method, provides broad coverage with high temporal resolution,

\* Corresponding author at: State Key Laboratory of Climate System Prediction and Risk Management (CPRM), School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing 210044, China.

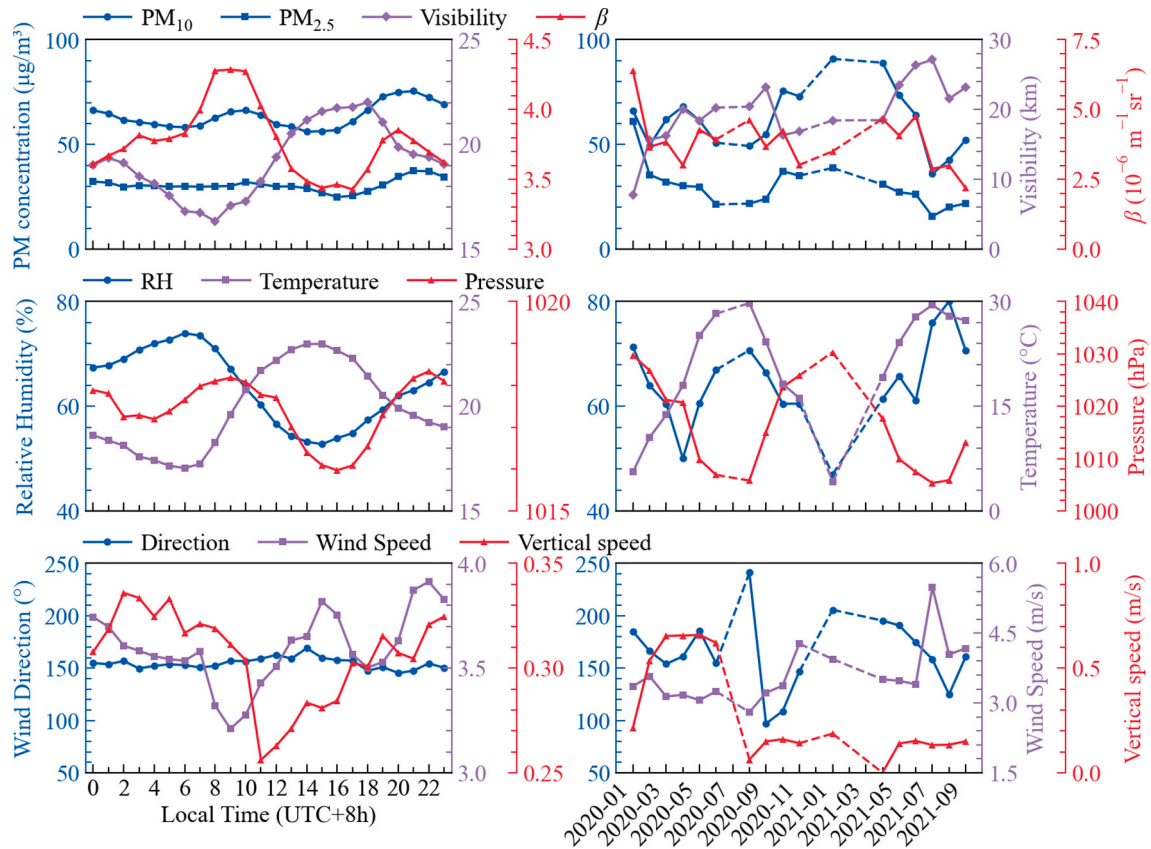
E-mail address: [twwei@nuist.edu.cn](mailto:twwei@nuist.edu.cn) (T. Wei).

<https://doi.org/10.1016/j.optlastec.2025.114338>

Received 29 July 2025; Received in revised form 14 November 2025; Accepted 18 November 2025

Available online 25 November 2025

0030-3992/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



**Fig. 1.** Hourly (left panels) and monthly (right panels) averaged results during 2020–2021: (a)  $PM_{10}$ ,  $PM_{2.5}$ , visibility and backscatter coefficient, (b) relative humidity, temperature and pressure, (c) horizontal wind direction, horizontal wind speed, and vertical wind speed. In the monthly series, dashed lines denote connections across non-consecutive months where data were unavailable.

yet faces limitations in retrieving sea fog horizontal visibility, especially during twilight or high cloud cover conditions [17,18,19], while PM retrievals are hindered by cloud interference, high-albedo surfaces, and severe pollution episodes [20,21]. Efforts to overcome these challenges increasingly leverage multi-source data fusion. Park et al. [22] integrated satellite-derived AOD, meteorological reanalysis, and land-use datasets within a convolutional neural network (CNN) framework, achieving enhanced  $PM_{2.5}$  estimation accuracy by resolving spatial heterogeneity in pollution drivers. In recent years, lidar (light detection and ranging) technology has been developed for detecting atmospheric components, aerosols, wind, cloud and precipitation [23,24,25,26,27]. Visibility retrieval relies on the lidar-based extinction measurement and a convert relation between the lidar wavelength and the wavelength of 550 nm that defines the visibility. For a simple single-wavelength aerosol lidar, the retrieval of the extinction coefficient is inherently linked to that of the backscattering coefficient, relying on an assumed lidar ratio. Horizontal measurements under homogeneous aerosol distribution have also been utilized to estimate independent extinction and horizontally average visibility [28,29]. PM estimation is more challenging due to its sensitivity to aerosol microphysical properties, including size distribution and hygroscopic growth [30,31]. Multi-wavelength lidars have been employed to retrieve aerosol size distributions and, consequently, PM concentration, typically using a regularization method [32,33]. However, the high cost and complexity of such systems limit their widespread application.

To overcome these challenges, recent advances integrate lidar observations with data-driven modeling. Shao et al. [34] developed a conversion model for retrieving  $PM_1$ ,  $PM_{2.5}$ , and  $PM_{10}$  concentrations from aerosol extinction coefficients, emphasizing the significant role of meteorological factors, particularly in the retrieval of  $PM_{10}$  concentrations. Zhen et al. [35] proposed a stacking fusion model that combines

XGBoost and LightGBM with the aim of improving atmospheric visibility prediction under varying pollution scenarios, using multivariate meteorological elements. These findings demonstrate the feasibility of applying neural networks to evaluate PM concentrations and visibility. Therefore, by integrating lidar-derived extinction coefficients with meteorological factors that influence the distribution of PM concentrations and visibility, and developing a neural network-based model to assess the spatial distribution of mass concentrations, it is possible to leverage the advantages of lidar in monitoring PM and visibility with wide coverage.

Coherent Doppler wind lidar is commonly used to measure wind and turbulence in the atmospheric boundary layer [36,37,38]. While backscatter intensity is often overlooked due to its sensitivity to heterodyne coupling efficiency, the underlying mechanism still relies on using aerosols as tracers to acquire Mie scattering signals. In recent years, Doppler lidar has been extended to retrieve aerosol backscatter and extinction coefficients [39,40]. Queißer et al. [41] demonstrated that continuous-wave Doppler wind lidar backscatter signals, when calibrated against co-located visibility sensors, provide a viable proxy for retrieving meteorological optical range (MOR), enabling remote and single-ended profiling of atmospheric visibility under vertically stratified aerosol conditions. With the deployment of numerous wind measurement lidars, the use of Doppler lidar backscatter intensity to retrieve visibility and PM concentrations will significantly broaden its application range.

In this study, we estimate visibility and PM concentrations by combining the backscatter coefficient of Doppler wind lidar with techniques such as linear regression and machine learning (ML). The results from this study demonstrate that lidar-based air quality monitoring networks can be further enhanced by incorporating additional meteorological data, beyond just the backscatter coefficient, thereby

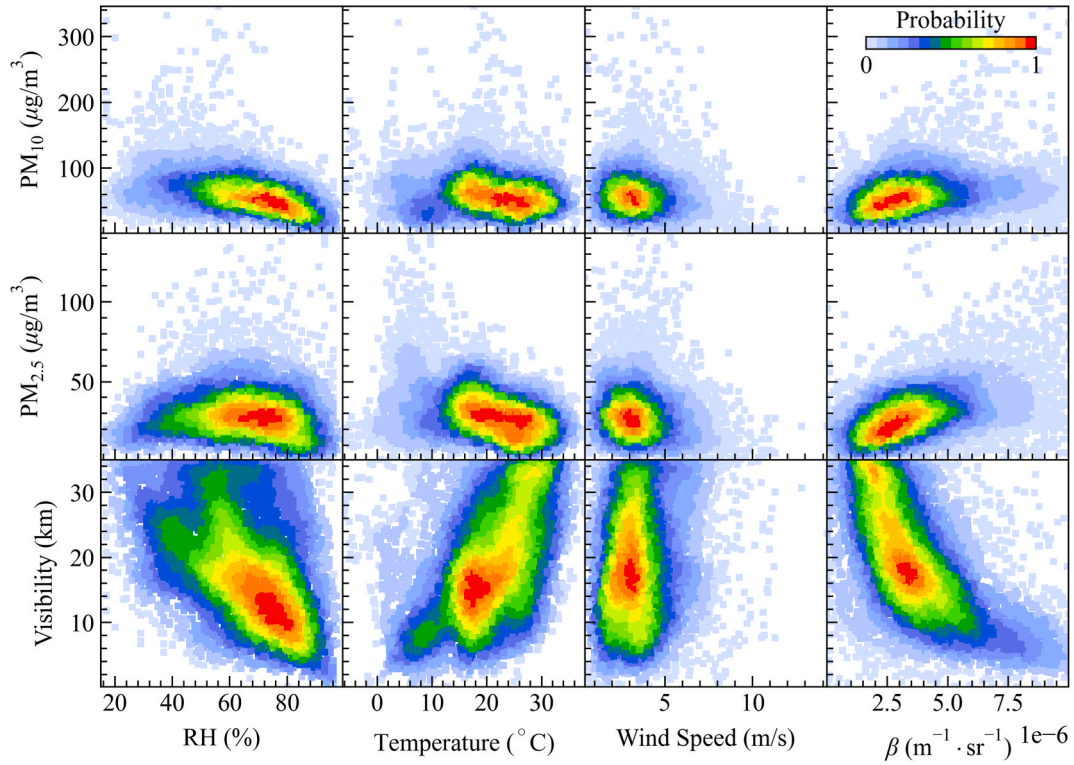


Fig. 2. Scatter plots of visibility,  $PM_{2.5}$  and  $PM_{10}$  versus relative humidity, temperature, horizontal wind speed and backscatter coefficient. The colorbar indicates the normalized density of data points.

improving predictive capabilities.

## 2. Data and method

### 2.1. Data and quality control

In this study, we employed a 1.5  $\mu\text{m}$  all-fiber coherent Doppler wind lidar to measure the backscatter coefficient  $\beta$ , horizontal wind speed, horizontal wind direction, and vertical wind speed. The data were collected from January 2020 to September 2021. The lidar system was deployed on the roof of the School of Earth and Space Science building of the University of Science and Technology of China (31.83° N, 117.25° E) in the urban area of Hefei, Anhui Province, China. The system operates with a laser wavelength of 1548 nm, emitting pulses of 300  $\mu\text{J}$  energy, 600 ns duration, and a repetition frequency of 10 kHz. The Doppler lidar employs a coaxial transmitter–receiver configuration and exhibits a near-field blind zone due to the specular reflection from the fiber end telescope. During the experiment, the system operated in VAD scan mode and the blind zone is approximately 60 m above ground level. We use the average of the lowest three valid range gates as the surface-layer proxy. Other detailed information about the lidar configuration, operation and data processing can be seen in previous studies [38,42,43,44]. An co-located automatic weather station supplied meteorological observations, including temperature, relative humidity, atmospheric pressure and visibility.  $PM_{2.5}$  and  $PM_{10}$  concentrations were obtained from the adjacent national air quality monitoring station. All measurements were pre-processed into 1-hour intervals. Data collected during precipitation events were excluded, and remaining missing values were imputed using k-nearest neighbors interpolation [45].

Fig. 1 shows the hourly and monthly mean trends of the dataset. Diurnal and seasonal variations can be seen clearly from most variables. The visibility exhibits inverse relationship with PM and backscatter coefficient. It tends to be lowest in the early morning, improves throughout the day, and reaches its peak around 18:00. Seasonal

patterns reveal lower visibility during winter and spring and higher values in summer. PM concentrations follow similar diurnal cycles, peaking at night. Furthermore, PM levels increase in spring and winter, primarily driven by Asian dust incursions and emissions from fossil fuels and industrial activities, resulting in higher PM concentrations [46]. Fig. 2 presents scatterplots of visibility,  $PM_{2.5}$  and  $PM_{10}$  against four key variables: RH, temperature, horizontal wind speed and  $\beta$ . In these plots, visibility generally decreases as  $\beta$  increase, while both  $PM_{2.5}$  and  $PM_{10}$  tend to rise under the same conditions. Lower temperatures are associated with higher  $PM_{2.5}$  levels and reduced visibility. Wind speed, however, shows no clear or systematic correlation with either visibility or particulate levels, which may be attributed to the complex interplay of wind-driven dispersion and pollutant transport processes [47].

When applying ML methods, all inputs were min–max normalized to ensure consistency and enhance model performance. To capture the inherent periodicity of temporal features [48], the month, day and hour variables were encoded via sine–cosine transformations:

$$\begin{cases} t\sin = \sin\left(\frac{2\pi \times t}{T}\right) \\ t\cos = \cos\left(\frac{2\pi \times t}{T}\right) \end{cases} \quad (1)$$

where  $t$  denotes the time component, and  $T$  its period (12 for months, 31 for days, and 24 for hours).

The machine learning methods used in this study optimized hyper-parameters using training and testing datasets, which comprised 7732 samples. In order to assess the performance, 5-fold cross-validation was applied with random sample-wise splits. The dataset was divided into five equal-sized, mutually exclusives folds, with one fold used for testing in each iteration and the remaining four folds used for training. The average performance metrics were computed across the five folds. The estimated visibility and PM concentrations were then compared with the observed data using Eqs. (2)–(5), and performance was evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE),



Mean Absolute Percentage Error (MAPE), and the correlation coefficient ( $R^2$ ).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{\text{est},i} - x_{\text{obs},i})^2} \quad (2)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_{\text{est},i} - x_{\text{obs},i}| \quad (3)$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{x_{\text{est},i} - x_{\text{obs},i}}{x_{\text{obs},i}} \right| \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_{\text{obs},i} - x_{\text{est},i})^2}{\sum_{i=1}^n (x_{\text{obs},i} - \bar{x}_{\text{obs}})^2} \quad (5)$$

## 2.2. Traditional regression methods

### 2.2.1. Visibility retrieval using regression methods

Atmospheric visibility is traditionally defined as the distance at which the intensity of a 550 nm light beam attenuates to 5 % of its original power under horizontally homogeneous atmospheric conditions. This distance is inversely proportional to the atmospheric extinction coefficient as expressed by Claus [49]

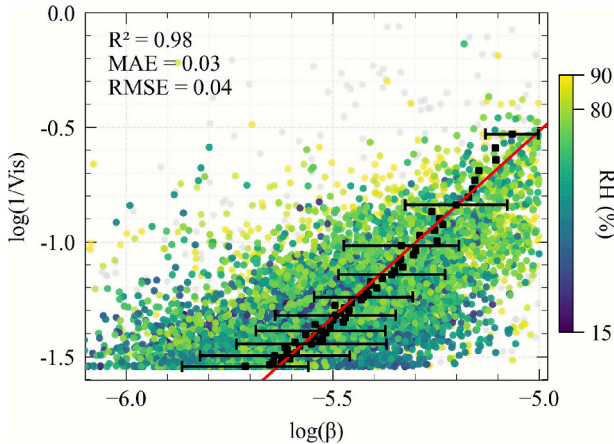
$$V = \frac{3.912}{\sigma} \quad (6)$$

Although the standard definition of visibility is based on 550 nm, lidar-based visibility detection typically employs alternative wavelengths due to practical considerations such as atmospheric transmission efficiency and cost-effectiveness. For example, 532 nm is widely used in visibility lidars [50,51]. Moreover, near-infrared wavelengths, such as 1.5  $\mu\text{m}$ , are increasingly employed for visibility detection, primarily due to its eye safety and the advantages of all-fiber architecture that facilitate system integration [28,52]. However, a transfer factor is required to convert the extinction coefficient at the measured wavelength to that at 550 nm based on the Ångström exponent.

In contrast to the physics-based approach described above, which is sensitive to uncertainties in the Ångström exponent and lidar ratio, an alternative statistical method establishes a direct empirical relationship between the backscatter coefficient and visibility. This method leverages a logarithmic-linear transfer function, expressed as [41]

$$\log_{10}(V^{-1}) = a \times \log_{10}(\beta) + b \quad (7)$$

This data-driven approach avoids aerosol-type assumptions and



**Fig. 3.** Scatter plot between inverse visibility and backscatter in logarithm unit. The black squares represent the binned mean values, with error bars indicating three times the standard deviation. The red line denotes the linear fit result.

directly correlates lidar measurements with visibility. Its primary advantage lies in its robustness and simplicity for long-term monitoring, effectively accommodating temporal variations in aerosol optical properties.

### 2.2.2. PM retrieval using regression methods

PM concentrations are influenced by hygroscopic growth, and the scattering enhancement factor  $f(RH)$ , quantifies how aerosol light-scattering coefficients vary with RH, reflecting the influence of chemical composition (e.g., sulfate, organic carbon) and particle size distribution [53]. Aerosols undergo deliquescence when RH surpasses the deliquescence relative humidity of substances like NaCl or  $(\text{NH}_4)_2\text{SO}_4$ , causing a rapid increase in scattering coefficients, depicted by a sharp rise in  $f(RH)$ . Conversely, efflorescence occurs at lower efflorescence relative humidity (ERH), resulting in a sudden reduction in  $f(RH)$  [54].

Based on the established relationships between aerosol backscatter and PM concentration, a regression model is formulated, incorporating the exponential dependence of backscatter on RH. This regression approach is grounded in empirical studies [54,55,56,57,58]:

$$\beta = k \times (\text{PM})^a \times (1 - \text{RH})^{-b \cdot \text{RH}} \quad (8)$$

where  $k$ ,  $a$ , and  $b$  are constants determined by least squares fitting, with PM expressed in units of mass concentration. For computational simplification, logarithmic transformation is applied to Eq. (8) before regression analysis.

## 2.3. ML-driven methods

This study employs three categories of supervised machine learning (ML) methods to estimate visibility and PM concentrations: tree-based models (Random Forest (RF), eXtreme Gradient Boosting (XGB), Light Gradient Boosting (LGB)), vector-based models (Support Vector Regression (SVR), and neural-based models (Multilayer Perceptron (MLP), Backpropagation Neural Network (BPNN), and Extreme Learning Machine (ELM)). For all models, the input features included the sine and cosine transformed temporal components (Month\_sin/cos, Day\_sin/cos, Hour\_sin/cos), along with meteorological and lidar-derived variables (Direction, RH, Pressure, Temperature,  $\beta$ , Horizontal wind speed, Vertical wind speed). The models were trained to predict three distinct output variables: Visibility,  $\text{PM}_{2.5}$ , and  $\text{PM}_{10}$ . Detailed information regarding the optimized hyperparameters for each model can be found in Table S1 in the Appendix.

### 2.3.1. Tree-based methods

Random Forest (RF) utilizes Bootstrap Aggregating (Bagging) and the Random Subspace Method (RSM) to generate multiple training subsets from random data samples [59]. Each subset trains a base learner, and predictions are averaged across all trees. In contrast, XGB and LGB employ boosting techniques for resampling and ensembling [60,61]. LGB enhances XGB by growing trees leaf-wise, accelerating learning and prediction, while XGB grows trees level-wise [62,63]. The predictions from each tree are weighed and averaged to obtain the result.

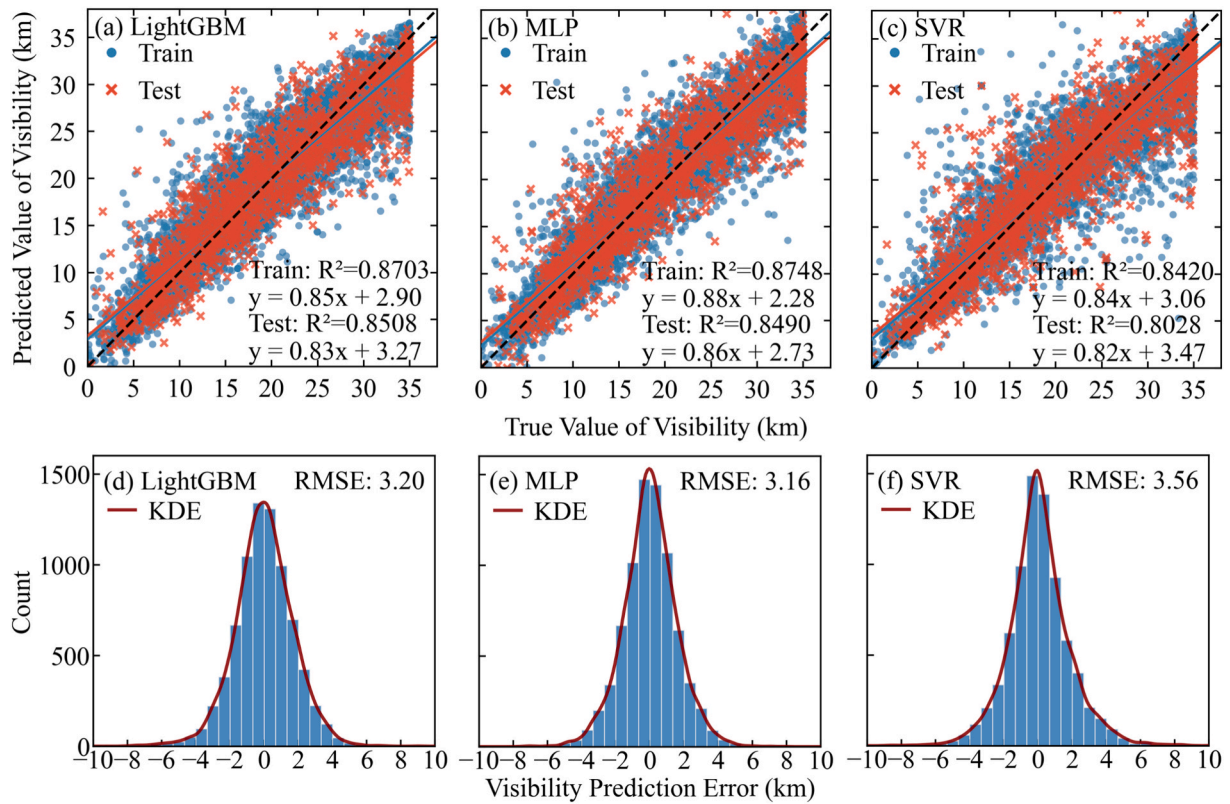
### 2.3.2. Vector-based methods

SVR, derived from the Support Vector Machine (SVM) framework [64], is a regression method that uses kernel functions to map non-linearly separable data into higher-dimensional spaces for linear separability [65]. The goal of SVR is to construct a regression model, where the loss occurs only when the absolute difference between the predicted value  $f(x)$  and the true value  $y$  exceeds a tolerance parameter  $\epsilon$ , creating a band of width  $2\epsilon$ .

### 2.3.3. Neural-based methods

MLP is a widely recognized type of artificial neural network,





**Fig. 4.** Comparison of ML models for visibility prediction. Panels (a–c) show scatter plots of predicted vs. observed visibility for (a) LGB, (b) MLP, and (c) SVR; blue points denote training data, red points denote test data, and the black dashed line indicates the 1:1 ideal ratio. Panels (d–f) show the corresponding error distributions for (d) LGB, (e) MLP, and (f) SVR, with red curves representing kernel density estimates.

extensively applied in classification and prediction tasks, particularly in domains such as image recognition [66]. Structurally, an MLP is composed of three primary layers: an input layer, one or more hidden layers, and an output layer. Based on the MLP architecture, the BPNN employs the backpropagation learning algorithm to iteratively adjust the weights of its input, hidden, and output layers through forward and backward passes, minimizing the error between predicted and actual outputs [67]. In contrast to the MLP model, the ELM model uses multiple perceptron to generate weights and biases both between the input and hidden layers and between the hidden and output layers to produce prediction results at the output layer [68].

**Table 1**

Five-fold cross-validated visibility prediction performance of LGB, MLP and SVR method.

| Model | Dataset | RMSE (km)       | MAE (km)        | MAPE (%)         | $R^2$               |
|-------|---------|-----------------|-----------------|------------------|---------------------|
| LGB   | Train   | $3.15 \pm 0.02$ | $2.32 \pm 0.01$ | $17.44 \pm 0.44$ | $0.8703 \pm 0.0013$ |
|       |         | $3.38 \pm 0.08$ | $2.48 \pm 0.05$ | $18.72 \pm 1.70$ | $0.8508 \pm 0.0061$ |
|       | Test    |                 |                 |                  |                     |
| MLP   | Train   | $3.10 \pm 0.06$ | $2.27 \pm 0.04$ | $17.03 \pm 0.70$ | $0.8748 \pm 0.0047$ |
|       |         | $3.40 \pm 0.09$ | $2.48 \pm 0.06$ | $18.79 \pm 3.53$ | $0.8490 \pm 0.0084$ |
|       | Test    |                 |                 |                  |                     |
| SVR   | Train   | $3.48 \pm 0.03$ | $2.23 \pm 0.01$ | $15.52 \pm 0.23$ | $0.8420 \pm 0.0021$ |
|       |         | $3.88 \pm 0.09$ | $2.67 \pm 0.04$ | $20.09 \pm 2.69$ | $0.8028 \pm 0.0085$ |
|       | Test    |                 |                 |                  |                     |

Note. Visibility is measured in kilometers (km); all values are reported as mean  $\pm$  standard deviation across five cross-validation folds.

### 3. Results

#### 3.1. Visibility estimation

##### 3.1.1. Traditional methods

Fig. 3 presents a scatter plot, depicting the logarithm of inverse visibility against the logarithm of lidar backscatter. The plot includes a color-coded RH distribution, reflecting aerosol hygroscopic growth characteristics: low-humidity ( $RH \leq 80\%$ , viridis color), mid-humidity ( $80\text{--}90\%$ , faded viridis color), and high-humidity ( $RH > 90\%$ , grey) conditions. In the low-humidity range, scatter points concentrate in the low region, showing an approximately linear trend where increasing corresponds to higher, indicating reduced visibility due to enhanced aerosol concentrations and optical attenuation [69]. Conversely, mid- and high-humidity conditions exhibit widely dispersed points without clear clustering, indicating a nonlinear relationship. This finding suggests that a linear visibility–backscatter relationship holds only within a limited range. The nonlinear behavior observed in these regions may result from rain-induced signal attenuation and the influence of low-level clouds or fog.

To investigate linear relationships in low-humidity regimes and mitigate fog/cloud contamination effects, we utilized the statistical calibration method of QueiBer et al. [41]. Visibility values below 3 km were excluded to ensure adherence to linear regression assumptions while reducing fog/cloud interference in visibility measurements. Aerosol backscatter coefficients were then calculated using visibility-stratified binning with 0.5 km intervals, deriving first-order moments ( $\mu_\beta$ , mean) and second-order moments ( $\delta_\beta$ , standard deviation). Subsequent linear regression analysis revealed strong agreement ( $R^2 = 0.98$ ), confirming the methodological robustness.

**Table 2**

MAE comparison across visibility ranges for Logarithmic-linear transfer, LGB, MLP, and SVR methods in predicting visibility.

| Model      | Dataset    | Visibility ranges (km) |                    |                     |                     |
|------------|------------|------------------------|--------------------|---------------------|---------------------|
|            |            | <5<br>(N = 264)        | 5–10<br>(N = 1050) | 10–25<br>(N = 4196) | 25–35<br>(N = 2222) |
| Log-Linear | RH < 80 %  | 2.64                   | 4.23               | 7.26                | 6.55                |
|            | Vis ≥ 3 km |                        |                    |                     |                     |
| LGB        | Train      | 1.94                   | 1.96               | 2.22                | 2.73                |
|            | Test       | 2.04                   | 2.10               | 2.33                | 2.93                |
| MLP        | Train      | 1.82                   | 1.91               | 2.14                | 2.74                |
|            | Test       | 2.13                   | 2.15               | 2.35                | 3.00                |
| SVR        | Train      | 1.50                   | 1.74               | 2.07                | 2.84                |
|            | Test       | 2.18                   | 2.25               | 2.47                | 3.31                |

Note. The number of samples in each range is indicated below the column headings.

### 3.1.2. ML-driven methods

To validate the effectiveness of ML methods, three representative models—LGB, MLP, and SVR—were selected for detailed analysis. These models were chosen for their distinctive methodological approaches: ensemble learning, neural networks, and vector-based regression, respectively. The performance metrics for the remaining models across both visibility and PM estimation tasks are provided in Table S2 in the Appendix.

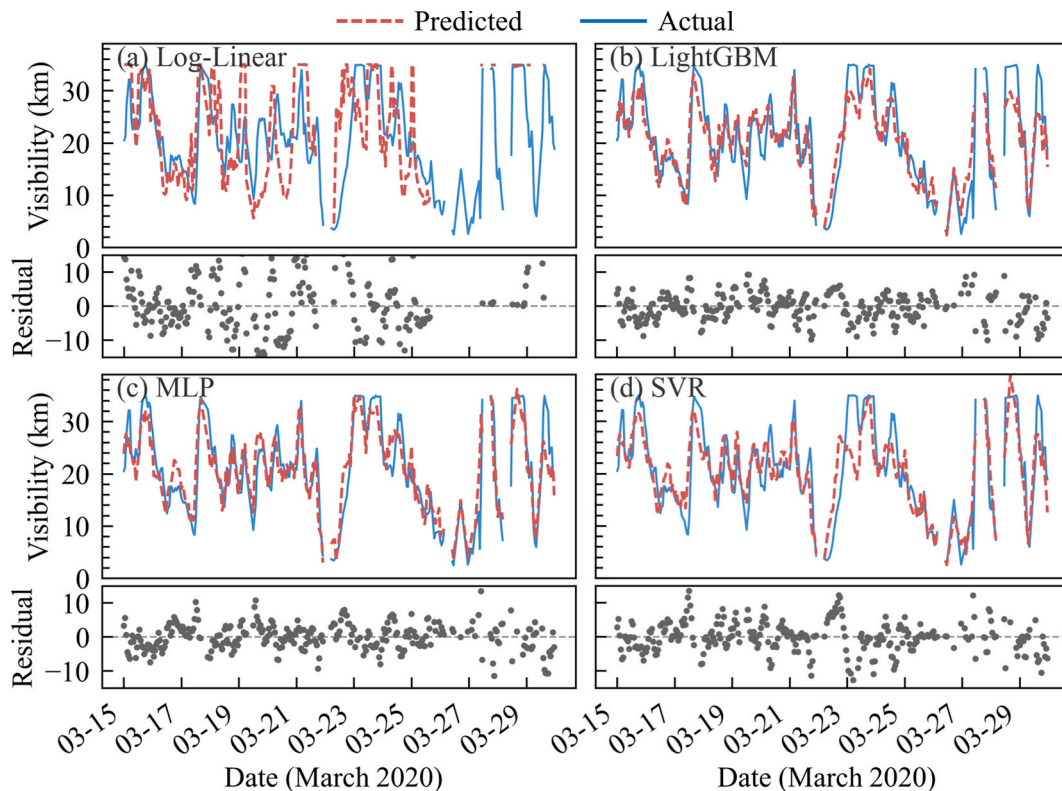
Fig. 4(a–c) shows the scatter plot between the estimated and true values for the three selected models. Among them, LGB achieves the best overall performance, with the lowest RMSE and highest  $R^2$ , benefiting from its leaf-wise growth strategy that efficiently captures complex nonlinear interactions [61,70]. MLP follows with slightly higher error metrics, while SVR trails behind due to its relatively rigid margin-based optimization, which limits its adaptability to varying aerosol conditions. To further evaluate the robustness of these models, a five-fold cross-validation was conducted, with results summarized in Table 1. The

results confirm that LGB provides the best overall performance and is the most robust model, consistently achieving the highest  $R^2$  and lowest error metrics on the test set. It also demonstrates the least performance fluctuation across the validation folds. Both MLP and SVR exhibit greater variability and lower accuracy compared to LGB.

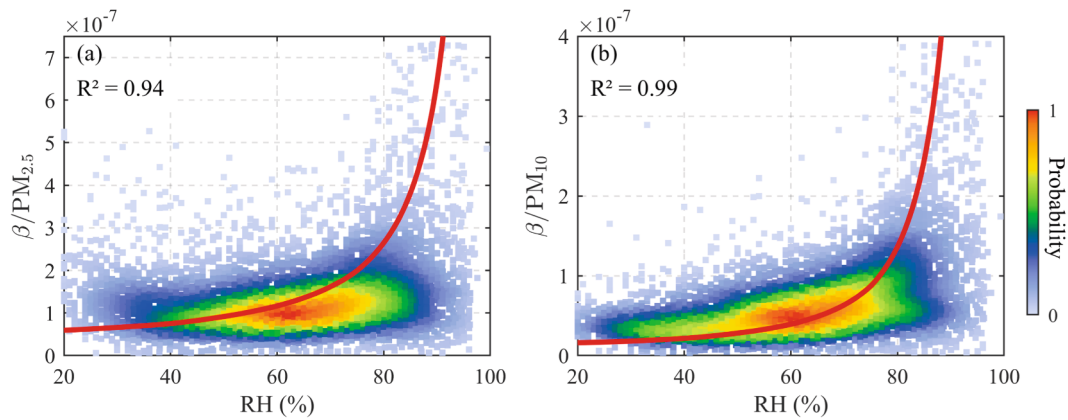
The error distributions for the three models, illustrated in Fig. 4(d–f), provide further insight into their predictive performance. All three models exhibit approximately normal residual distributions centered near zero, suggesting a lack of significant systematic bias in their predictions. Notably, the error spreads are comparable across the models, with the majority of residuals for each falling within approximately  $\pm 6$  km. The overall RMSE for all data points indicates that MLP achieves the lowest error (3.16 km), closely followed by LGB (3.20 km). Both slightly outperform SVR, which has a higher overall RMSE of 3.56 km. These results suggest that both MLP and LGB deliver a superior and highly comparable level of predictive accuracy.

Table 2 compares the MAE of the three ML models against the univariate logarithmic-linear transfer method across different visibility bands. The results clearly demonstrate that all ML methods consistently outperform the traditional approach in every tested range. Among the ML models, LGB establishes itself as the most accurate model, achieving the lowest test-set MAE across all four visibility bands. While MLP also delivers highly competitive performance, its MAE is consistently slightly higher than that of LGB, whereas SVR trails both models. This detailed analysis confirms that LGB provides superior visibility estimation, attributed to its effective regularization and capacity to model intricate nonlinear relationships.

While the MAE for all ML methods is numerically lower in the <5 km visibility range (Table 2), this metric can be misleading when viewed in isolation. To provide a more rigorous assessment of the model's explanatory power, we calculated the  $R^2$ . For this specific range, the LGB's  $R^2$  value was only 0.062. This degradation in performance is likely due to the scarcity of sub-5 km samples in our dataset. Fig. 5 presents a time-series comparison of observed and predicted visibility between



**Fig. 5.** Time-series and residual plots comparison of methods for visibility prediction from 2020 to 03-15 to 2020-03-30 with precipitation periods excluded: (a) logarithmic-linear transfer applied to the RH < 80 % subset, (b) LGB, (c) MLP, (d) SVR.



**Fig. 6.** Scatter plot of  $\beta PM$  versus RH, overlaid with the multiple-regression fit from Section 2.2.2. The color represents normalized data-point density. The  $R^2$  values shown are from the regression fit to binned statistical data points.

**Table 3**

Five-fold cross-validated PM prediction performance of LGB, MLP and SVR method.

| Setting           | Model            | Dataset | RMSE<br>( $\mu\text{g}/\text{m}^3$ ) | MAE<br>( $\mu\text{g}/\text{m}^3$ ) | MAPE<br>(%)      | $R^2$               |
|-------------------|------------------|---------|--------------------------------------|-------------------------------------|------------------|---------------------|
| PM <sub>2.5</sub> | LGB              | Train   | 5.92 $\pm$ 0.05                      | 4.24 $\pm$ 0.02                     | 19.91 $\pm$ 0.29 | 0.8758 $\pm$ 0.0008 |
|                   |                  | Test    | 7.02 $\pm$ 0.37                      | 4.89 $\pm$ 0.16                     | 23.05 $\pm$ 1.84 | 0.8254 $\pm$ 0.0111 |
|                   | MLP              | Train   | 6.71 $\pm$ 0.11                      | 4.88 $\pm$ 0.11                     | 22.44 $\pm$ 0.47 | 0.8405 $\pm$ 0.0053 |
|                   |                  | Test    | 7.48 $\pm$ 0.35                      | 5.36 $\pm$ 0.19                     | 24.81 $\pm$ 1.56 | 0.8016 $\pm$ 0.0130 |
|                   | SVR              | Train   | 7.77 $\pm$ 0.10                      | 4.87 $\pm$ 0.04                     | 21.91 $\pm$ 0.57 | 0.7861 $\pm$ 0.0043 |
|                   |                  | Test    | 8.67 $\pm$ 0.34                      | 5.87 $\pm$ 0.14                     | 26.89 $\pm$ 2.68 | 0.7332 $\pm$ 0.0175 |
|                   | PM <sub>10</sub> | LGB     | 11.33 $\pm$ 0.21                     | 7.80 $\pm$ 0.12                     | 16.63 $\pm$ 0.16 | 0.8944 $\pm$ 0.0026 |
|                   |                  |         | 15.32 $\pm$ 0.91                     | 10.33 $\pm$ 0.21                    | 22.20 $\pm$ 1.11 | 0.8063 $\pm$ 0.0143 |
|                   |                  | MLP     | 12.57 $\pm$ 0.93                     | 9.01 $\pm$ 0.66                     | 19.02 $\pm$ 1.47 | 0.8693 $\pm$ 0.0198 |
|                   |                  |         | 15.84 $\pm$ 0.61                     | 11.12 $\pm$ 0.30                    | 23.38 $\pm$ 1.06 | 0.7927 $\pm$ 0.0104 |
|                   |                  | SVR     | 17.86 $\pm$ 0.28                     | 10.16 $\pm$ 0.08                    | 19.25 $\pm$ 0.33 | 0.7376 $\pm$ 0.0038 |
|                   |                  |         | 19.04 $\pm$ 1.58                     | 11.68 $\pm$ 0.38                    | 23.15 $\pm$ 1.90 | 0.7014 $\pm$ 0.0232 |

March 15–30, 2020. Noticeable discrepancies occur at higher visibility levels, with deviations between predictions and observations becoming more pronounced in the 25–35 km range. In contrast, for moderate visibility conditions (10–25 km), the predictions align more closely with observations, demonstrating stable model performance. Collectively, these results indicate that the model's predictive performance requires improvement under both low and high extreme visibility conditions.

### 3.2. PM estimation

#### 3.2.1. Traditional methods

To characterize the humidity dependence of aerosol backscatter, RH values were first binned at 2 % intervals, and within each bin we computed the median  $\beta/PM$ ,  $PM_{2.5}$ , and  $PM_{10}$ . These aggregated data points were then subject to the multiple-regression scheme described in Section 2.2.2. Fig. 6 presents the resulting scatter of  $\beta/PM$  versus RH, overlaid with the fitted regression curve. Notably,  $\beta/PM$  exhibits an exponential relationship with RH: it rises slowly when RH is below 80 %,

but increases sharply once RH exceeds 80 %, reflecting aerosol deliquescence. This behavior aligns closely with the findings of Zieger et al. [71], Won et al. [72]. However, notable differences in fitting performance are observed between  $PM_{2.5}$  and  $PM_{10}$ . The  $PM_{10}$  yields a higher coefficient of determination than  $PM_{2.5}$ , as evidenced by its regression curve more closely following the high-data-density regions, indicating a superior model fit for coarse-mode aerosols. This can be attributed to the 1.5  $\mu\text{m}$  lidar wavelength used in this study, which is more sensitive to larger particles compared with traditional 905 nm ceilometers [73]. In addition, the  $\beta/PM_{2.5}$  exhibits a much stronger increase at high RH values, suggesting that coarse-mode aerosols contribute more significantly to light scattering through hygroscopic growth.

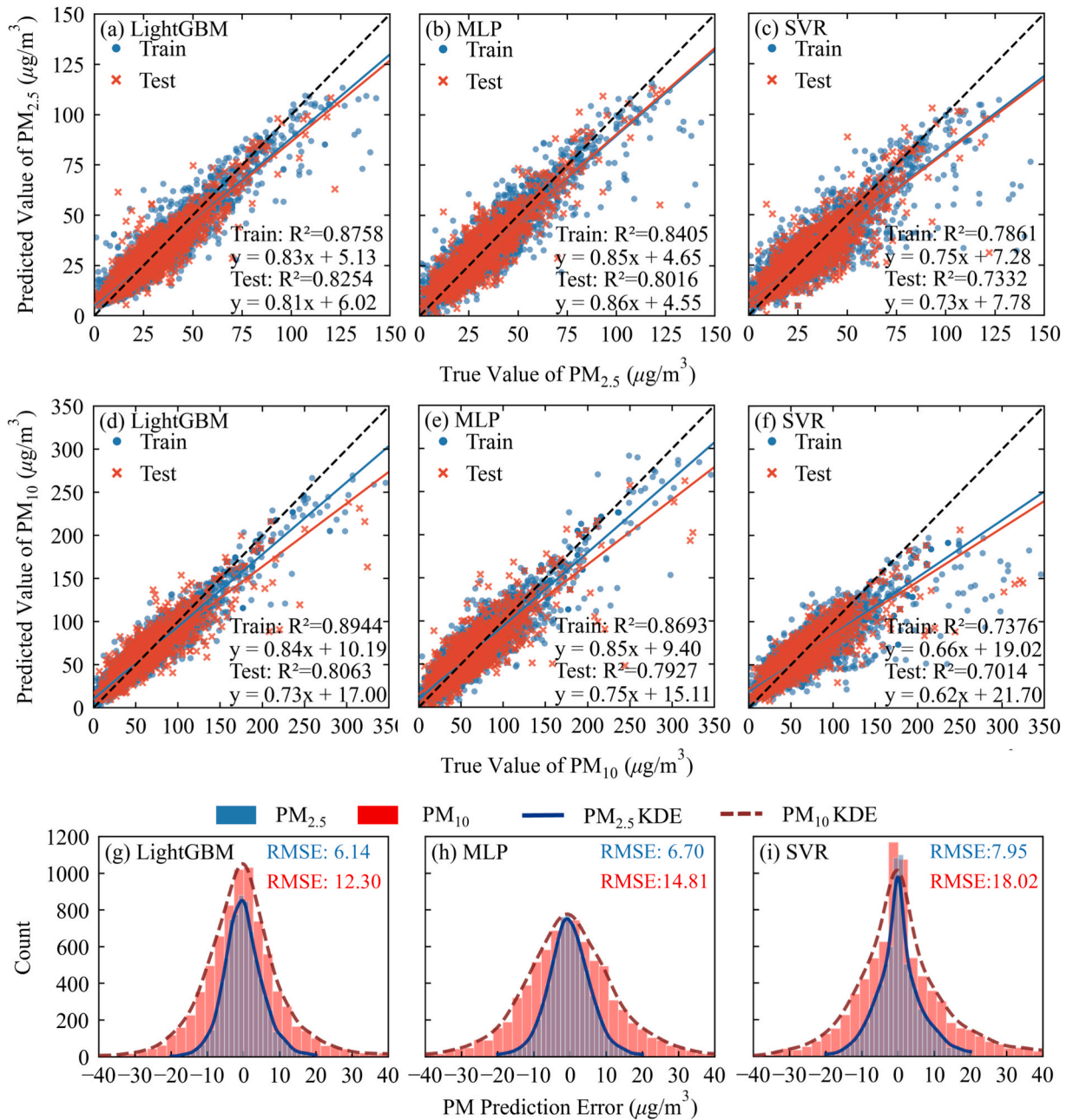
#### 3.2.2. ML-driven methods

To benchmark against the ML methods, the regression equation fitted in Section 3.2.1 was applied to the full dataset to compute PM concentrations for every combination of RH and backscatter coefficient. Table 3 presents the five-fold cross-validated performance metrics for  $PM_{2.5}$  and  $PM_{10}$  prediction. Fig. 7's scatter plots further demonstrate that LGB's  $PM_{2.5}$  and  $PM_{10}$  predictions exhibit the fewest deviations from the 1:1 ideal line, underscoring its superior accuracy among the three methods.

The error distributions in Fig. 7(g-i) present a complex view. While SVR shows the highest concentration of residuals exactly at zero, its error distribution is also the widest, indicating a higher frequency of significant errors. A more definitive evaluation using the overall RMSE across all data confirms the superior performance of LGB. For  $PM_{2.5}$  and  $PM_{10}$ , LGB achieves the lowest RMSEs (6.14 and 12.30  $\mu\text{g}/\text{m}^3$ ), clearly outperforming MLP (6.70 and 14.81  $\mu\text{g}/\text{m}^3$ ) and SVR (7.95 and 18.02  $\mu\text{g}/\text{m}^3$ ). This quantitative analysis solidifies LGB's position as the most accurate and reliable model for PM estimation, as its predictions have the lowest overall error.

Table 4 compares the MAE for  $PM_{2.5}$  and  $PM_{10}$  estimations across true concentration intervals using the multiple-regression scheme and the three ML methods. The ML models consistently outperform the multiple-regression approach, which incurs high MAEs of 10.08–63.30  $\mu\text{g}/\text{m}^3$  for  $PM_{2.5}$  and 20.87–54.57  $\mu\text{g}/\text{m}^3$  for  $PM_{10}$ . Among the ML models, LGB delivered the most dominant performance for  $PM_{2.5}$ , achieving the lowest test-set MAE in every concentration bin. The results for  $PM_{10}$  were more varied: SVR performed best at low concentrations (<50  $\mu\text{g}/\text{m}^3$ ), MLP was most accurate at the highest concentrations (>150  $\mu\text{g}/\text{m}^3$ ), and LGB excelled in the intermediate ranges. Fig. 8's time-series plots further illustrate this heterogeneous model performance and confirm that estimation errors are larger at high PM concentrations, as evidenced by the performance of LGB and SVR. Consistent with the findings for low-visibility conditions, this is likely due to the limited number of high-pollution samples in the dataset. In





**Fig. 7.** Comparison of ML models for PM prediction. Panels (a–c) show scatter plots of predicted vs. observed PM<sub>2.5</sub> for (a) LGB, (b) MLP, and (c) SVR; panels (d–f) show scatter plots for PM<sub>10</sub> for (d) LGB, (e) MLP, and (f) SVR. Blue points denote training data, red points denote test data, and the black dashed line indicates the 1:1 ideal ratio. Panels (g–i) present the corresponding error distribution histograms for (g) LGB, (h) MLP, and (i) SVR, with red curves representing kernel density estimates.

summary, all ML-based models markedly outperform the multiple-regression scheme, with LGB delivering the most consistent gains in accuracy and robustness, particularly for PM<sub>2.5</sub> estimation.

### 3.3. Feature importance in ML-based estimation

To better understand the mechanisms driving model performance, we analyzed the relative importance of input features in the LGB models using the 'gain' metric, which quantifies the improvement in the model's objective function from each feature. The results, shown in Fig. 9, provide insight into the physical and statistical relevance of the meteorological and lidar-derived variables for each estimation task.

For visibility estimation,  $\beta$  remains the most influential feature, contributing 39.42 % to the model. This aligns with the physical

principle that visibility is primarily governed by aerosol scattering and absorption, which are directly related to the backscatter signal [3]. RH emerges as the second most important predictor (18.43 %), followed closely by temperature (15.33 %). High RH enhances light scattering through aerosol hygroscopic growth, while temperature can influence atmospheric stability and the formation rate of secondary aerosols, both of which impact visibility [74,75].

For PM<sub>2.5</sub> concentration estimation,  $\beta$  is the most influential feature at 23.25 %, indicating the model successfully learned a direct relationship between the optical properties measured by the lidar and the mass of fine particulate matter. The model also heavily depends on temporal and meteorological inputs. The seasonal cycle, represented by Month<sub>-</sub>cos (14.81 %), is the second most important feature, highlighting the strong seasonal variations in PM<sub>2.5</sub> driven by factors like winter heating

**Table 4**

MAE comparison across PM ranges for Multiple-Regression, LGB, MLP, and SVR methods in predicting PM Concentrations (Unit:  $\mu\text{g}/\text{m}^3$ ).

| Setting           | Model               | Dataset | PM <sub>2.5</sub> concentrations ranges ( $\mu\text{g}/\text{m}^3$ ) |                      |                      |                   |
|-------------------|---------------------|---------|--|----------------------|----------------------|-------------------|
|                   |                     |         | <20<br>(N = 2125)  | 20–60<br>(N = 5202)  | 60–100<br>(N = 329)  | >100<br>(N = 76)  |
| PM <sub>2.5</sub> | Multiple-Regression | All     | 10.08  | 11.97                | 27.17                | 63.30             |
|                   | LGB                 | Train   | 3.92   | 4.04                 | 7.12                 | 18.01             |
|                   |                     | Test    | 4.49   | 4.60                 | 8.85                 | 23.16             |
|                   | MLP                 | Train   | 4.38   | 4.72                 | 7.99                 | 20.58             |
|                   |                     | Test    | 4.73   | 5.18                 | 9.02                 | 23.71             |
|                   | SVR                 | Train   | 4.32   | 4.48                 | 10.09                | 31.49             |
|                   |                     | Test    | 5.37   | 5.38                 | 11.94                | 34.36             |
| Setting           | Model               | Dataset | PM <sub>10</sub> concentrations ranges ( $\mu\text{g}/\text{m}^3$ )  |                      |                      |                   |
|                   |                     |         | <50<br>(N = 2860)  | 50–100<br>(N = 4003) | 100–150<br>(N = 667) | >150<br>(N = 202) |
| PM <sub>10</sub>  | Multiple-Regression | All     | 20.87  | 27.95                | 31.41                | 54.57             |
|                   | LGB                 | Train   | 6.78   | 6.87                 | 13.35                | 22.55             |
|                   |                     | Test    | 8.91   | 8.91                 | 17.90                | 33.30             |
|                   | MLP                 | Train   | 8.64   | 9.26                 | 15.59                | 24.04             |
|                   |                     | Test    | 10.58  | 11.04                | 18.73                | 31.82             |
|                   | SVR                 | Train   | 7.04   | 8.33                 | 20.03                | 58.18             |
|                   |                     | Test    | 8.59   | 9.69                 | 21.95                | 60.68             |

Note. The number of samples in each range is indicated below the column headings.

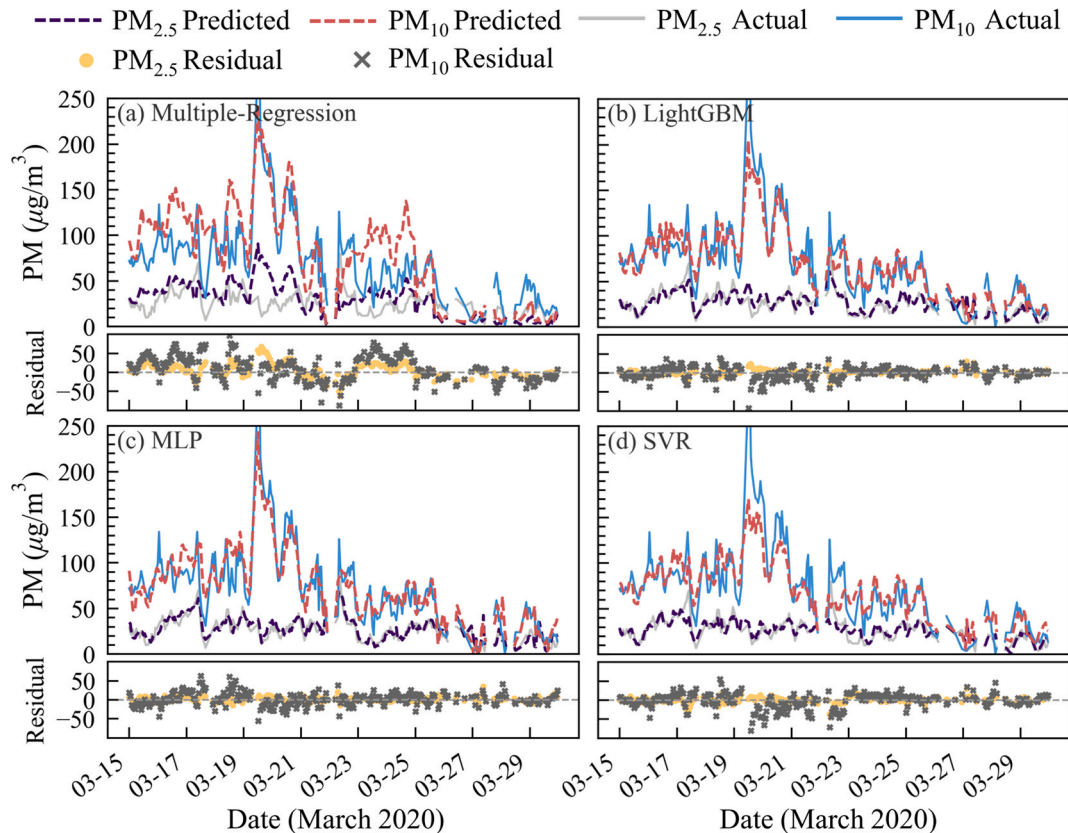
and summertime photochemistry [76,77]. Atmospheric pressure (11.86 %) and RH (10.24 %) are also critical, as high-pressure systems often lead to stable conditions that trap pollutants, and humidity is crucial for secondary aerosol formation processes [78,79,80].

In the case of PM<sub>10</sub> estimation, RH is the most dominant feature (26.19 %) followed by  $\beta$  (18.92 %). This is consistent with our earlier observation that  $\beta/\text{PM}_{10}$  exhibits a stronger increase at high RH values compared to  $\beta/\text{PM}_{2.5}$ , suggesting that coarse-mode aerosols contribute more significantly to light scattering through hygroscopic growth. Additionally, high humidity often indicates increased atmospheric stability and stagnant air, which inhibit the dispersion of pollutants and lead to the accumulation of PM<sub>10</sub> [81]. Furthermore, the model identified vertical wind speed as a more significant predictor for PM<sub>10</sub> than for PM<sub>2.5</sub>. This distinction underscores the different physical dynamics of coarse versus fine particles. The PM<sub>10</sub> fraction often includes mechanically generated particles that are more susceptible to gravitational settling. Consequently, vertical air motions play a more critical role in either maintaining the suspension of these larger particles or accelerating their deposition [82]. The longer atmospheric residence time of fine PM<sub>2.5</sub> particles makes their concentrations comparatively less sensitive to such immediate vertical wind fluctuations [1].

#### 4. Discussion and conclusion

This study demonstrates the potential of Doppler wind lidar, when combined with meteorological parameters, for estimating visibility and particulate matter concentrations using both traditional and machine learning approaches. Among all methods tested, ensemble tree-based ML models, particularly Light Gradient Boosting, consistently outperformed conventional regression techniques in atmospheric parameter estimation.

Quantitative assessment reveals that LGB achieves high accuracy in



**Fig. 8.** Time-series and residual plots comparison of methods for PM prediction from 2020-03-15 to 2020-03-30 with precipitation periods excluded: (a) LGB, (b) MLP, (c) SVR, (d) Multiple-Regression

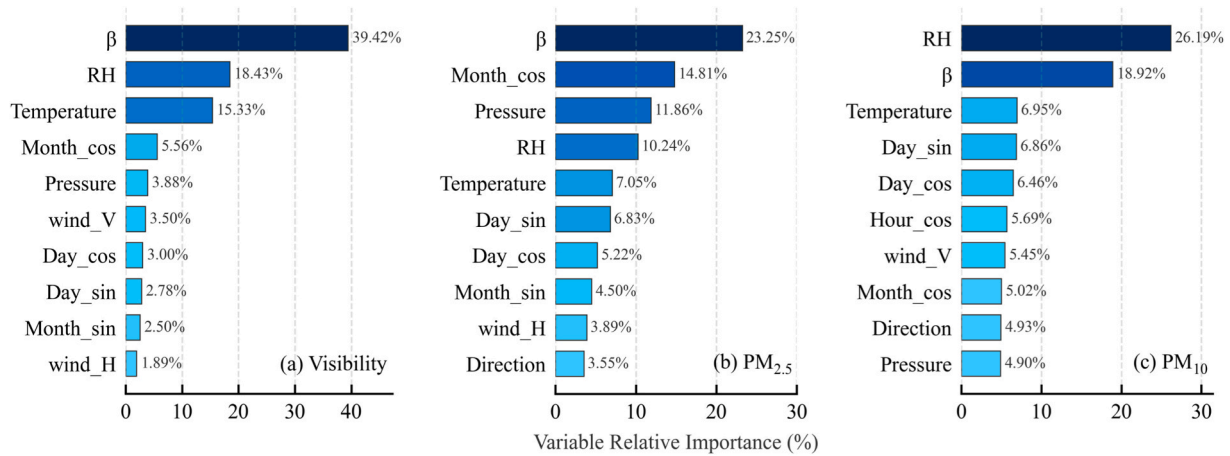


Fig. 9. Top 10 most important input variables for estimation using LGB: (a) Visibility (b) PM<sub>2.5</sub> (c) PM<sub>10</sub>.

visibility prediction, with a test-set RMSE of  $3.38 \pm 0.08$  km and  $R^2$  of  $0.8508 \pm 0.0061$ , substantially exceeding the capabilities of traditional logarithmic-linear transfer models. Similarly, for PM concentration estimation, LGB yielded RMSE values of  $7.02 \pm 0.37$   $\mu\text{g}/\text{m}^3$  ( $R^2 = 0.8254 \pm 0.0111$ ) for PM<sub>2.5</sub>, and  $15.32 \pm 0.91$   $\mu\text{g}/\text{m}^3$  ( $R^2 = 0.8063 \pm 0.0143$ ) for PM<sub>10</sub>, outperforming conventional multiple regression approaches.

Feature importance analysis reveals fundamental differences in the physical mechanisms governing visibility and PM estimation. The backscatter coefficient dominates visibility prediction (39.42 %), confirming that visibility is primarily determined by light scattering and extinction processes directly measured through lidar backscatter. PM estimation shows distinct patterns across particle sizes. For PM<sub>2.5</sub>, the backscatter coefficient is most important (23.25 %), with seasonal cycles (Month\_cos: 14.81 %) and atmospheric pressure (11.86 %) also contributing significantly, reflecting seasonal pollution patterns and meteorological trapping effects. For PM<sub>10</sub>, RH dominates (26.19 %) as a proxy for atmospheric stability conditions that promote coarse particle accumulation, while β ranks second (18.92 %). Notably, vertical wind speed is more important for PM<sub>10</sub> than PM<sub>2.5</sub>, reflecting the greater susceptibility of coarse particles to gravitational settling and vertical transport processes.

Despite the promising results, several limitations warrant consideration. First, the model's predictive accuracy diminishes under low-visibility conditions and high PM concentration scenarios, highlighting challenges in characterizing extreme events with limited representation in the training dataset. Second, our approach primarily utilizes near-surface lidar measurements, potentially overlooking the vertical distribution of aerosols that may be crucial for understanding complex pollution conditions. Third, since the models were developed and validated using data from a specific location in Hefei, China, their transferability to regions with different aerosol characteristics and meteorological conditions requires further investigation. While this study demonstrates the potential of Doppler wind lidar for visibility and PM estimation at a single site, further research is needed to validate the method across multiple sites and broader spatial scales. Fourth, while we employed separate models for each target variable to ensure practical applicability and enable fair comparison with traditional models, we also explored a multi-task learning approach that jointly predicts all three variables simultaneously. Drawing inspiration from recent advances in multi-task learning for air quality estimation [83], we developed and evaluated a multi-task LGB formulation to leverage the inherent correlations among these atmospheric parameters. The multi-task results, presented in Table S3 of the supplementary materials, demonstrate comparable performance to individual models while offering computational efficiency advantages. Specifically, the multi-task

model yielded slightly higher test RMSEs, with degradation of +1.8 % for visibility (from 3.38 to 3.44 km), +6.1 % for PM<sub>2.5</sub> (from 7.02 to 7.45  $\mu\text{g}/\text{m}^3$ ), and +7.8 % for PM<sub>10</sub> (from 15.32 to 16.52  $\mu\text{g}/\text{m}^3$ ) compared to the optimized single-task models. This minor reduction in accuracy is balanced by the significant computational benefit of training a single model instead of three, making the approach highly suitable for operational applications.

In conclusion, this study demonstrates the feasibility and effectiveness of integrating Doppler wind lidar data with meteorological parameters through machine learning techniques to achieve high-accuracy estimation of visibility and PM concentrations. By repurposing existing lidar infrastructure and implementing advanced data fusion methods, we have developed a promising approach that addresses the limitations of conventional point-based measurement systems. The methodology established herein offers significant potential for practical applications in air quality management, transportation safety, and urban planning, particularly in regions with limited monitoring resources. Future research will focus on incorporating vertical profile information, expanding the training dataset to include more extreme events, validating the methodology across diverse atmospheric conditions, and exploring additional data sources to further enhance retrieval capabilities.

#### CRedit authorship contribution statement

**Wenhan Xu:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Tianwen Wei:** Writing – review & editing, Supervision, Resources, Funding acquisition, Data curation, Conceptualization. **Mengya Wang:** Writing – review & editing, Validation, Resources. **Zhekai Li:** Writing – review & editing, Validation. **Zhuoqun Li:** Validation, Investigation. **Zhen Zhang:** Validation, Supervision. **Haiyun Xia:** Writing – review & editing, Supervision, Resources.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

Thanks for the support of the National Key Laboratory of Climate System Prediction and Change Response (CPRM-2025-NUIS-012), the National Natural Science Foundation of China (42405136), the China Meteorological Administration Xiong'an Atmospheric Boundary Layer Key Laboratory (2023LABL-B11).



## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.optlastec.2025.114338>.

## Data availability

Data will be made available on request.

## References

- [1] U. Pöschl, Atmospheric aerosols: composition, transformation, climate and health effects, *Angew. Chem. Int. Ed.* 44 (2005) 7520–7540, <https://doi.org/10.1002/chin.200607299>.
- [2] Z. Chen, J.-N. Wang, G.-X. Ma, Y.-S. Zhang, China tackles the health effects of air pollution, *Lancet* 382 (2013) 1959–1960, [https://doi.org/10.1016/s0140-6736\(13\)62064-4](https://doi.org/10.1016/s0140-6736(13)62064-4).
- [3] S. Han, et al., Effect of aerosols on visibility and radiation in spring 2009 in Tianjin, China, *Aerosol Air Qual. Res.* 12 (2012) 211–217, <https://doi.org/10.4209/aaqr.2011.05.0073>.
- [4] R. Zhang, et al., Formation of urban fine particulate matter, *Chem. Rev.* 115 (2015) 3803–3855, <https://doi.org/10.1021/acs.chemrev.5b00067>.
- [5] WMO, 2014: Guide to meteorological instruments and methods of observation (WMO-No. 8). Geneva: World Meteorological Organization, 29.
- [6] M. Woody, H.-W. Wong, J. West, S. Arunachalam, Multiscale predictions of aviation-attributable PM<sub>2.5</sub> for US airports modeled using CMAQ with plume-in-grid and an aircraft-specific 1-D emission model, *Atmos. Environ.* 147 (2016) 384–394, <https://doi.org/10.1016/j.atmosenv.2016.10.016>.
- [7] W.-S. Won, R. Oh, W. Lee, K.-Y. Kim, S. Ku, P.-C. Su, Y.-J. Yoon, Impact of fine particulate matter on visibility at Incheon International Airport, South Korea, *Aerosol Air Quality Research* 20 (2020) 1048–1061, <https://doi.org/10.4209/aaqr.2019.03.0106>.
- [8] X. Liu, et al., Increase of aerosol scattering by hygroscopic growth: Observation, modeling, and implications on visibility, *Atmos. Res.* 132 (2013) 91–101, <https://doi.org/10.1016/j.atmosres.2013.04.007>.
- [9] W. Chen, et al., Chemical composition of PM<sub>2.5</sub> and its impact on visibility in Guangzhou, Southern China, *Aerosol Air Qual. Res.* 16 (2016) 2349–2361, <https://doi.org/10.4209/aaqr.2016.02.0059>.
- [10] J. Wang, et al., Exploring the sensitivity of visibility to PM<sub>2.5</sub> mass concentration and relative humidity for different aerosol types, *Atmos. Res.* 13 (2022) 471, <https://doi.org/10.3390/atmos13030471>.
- [11] Li, K. H., N. D. Le, L. Sun, and J. V. Zidek, 1999: Spatial-temporal models for ambient hourly PM<sub>10</sub> in Vancouver. *Environmetrics: The official journal of the International Environmetrics Society*, 10, 321–338, DOI: 10.1002/(sici)1099-095x(199905/06)10:3<321::aid-env355>3.0.co;2-d.
- [12] J. Van der Wal, L. Janssen, Analysis of spatial and temporal variations of PM<sub>10</sub> concentrations in the Netherlands using Kalman filtering, *Atmos. Environ.* 34 (2000) 3675–3687, [https://doi.org/10.1016/s1352-2310\(00\)00085-6](https://doi.org/10.1016/s1352-2310(00)00085-6).
- [13] M. Fu, F. Zheng, X. Xu, L. Niu, Advances of study on monitoring and evaluation of PM<sub>2.5</sub> pollution, *Meteorol. Disaster Reduc. Res.* 34 (2011) 1–6, <https://doi.org/10.2139/ssrn.5103722>.
- [14] C. Mazzoleni, H.D. Kuhns, H. Moosmüller, Monitoring automotive particulate matter emissions with lidar: a review, *Remote Sens. (Basel)* 2 (2010) 1077–1119, <https://doi.org/10.3390/rs2041077>.
- [15] X. Xing, Y. Cui, F. Zhang, B. Xie, Summary of present situation and development trend of visibility measurement technology, *Metrol. Meas. Technol.* 30 (2010) 15–20, <https://doi.org/10.3969/j.issn.1674-5795.2010.05.005>, in Chinese.
- [16] H. Merbitz, S. Fritz, C. Schneider, Mobile measurements and regression modeling of the spatial particulate matter variability in an urban area, *Sci. Total Environ.* 438 (2012) 389–403, <https://doi.org/10.1016/j.scitotenv.2012.08.049>.
- [17] J.-M. Yoo, M.-J. Jeong, Y.M. Hur, D.-B. Shin, Improved fog detection from satellite in the presence of clouds, *Asia-Pac. J. Atmos. Sci.* 46 (2010) 29–40, <https://doi.org/10.1007/s13143-010-0004-5>.
- [18] E.M. Wilcox, Multi-spectral remote sensing of sea fog with simultaneous passive infrared and microwave sensors, *Marine Fog: Challenges Advancements in Observations, Modeling, Forecasting* 511–526 (2017), [https://doi.org/10.1007/978-3-319-45229-6\\_11](https://doi.org/10.1007/978-3-319-45229-6_11).
- [19] Z. Yang, M. Wu, M. Xu, X. Zhu, C. Zhang, B. Zhang, MoANet: a Motion attention Network for Sea Fog Detection in Time Series Meteorological Satellite Imagery, *IEEE J. Selected Top. Appl. Earth Observ. Remote Sens.* 17 (2023) 1976–1987, <https://doi.org/10.1109/jstars.2023.3340909>.
- [20] J. Pelon, D.M. Winker, G. Ancellet, M.A. Vaughan, D. Josset, A. Bazureau, N. Pascal, Space observation of aerosols from satellite over China during pollution episodes: status and perspectives, *Air Pollut. Eastern Asia Integrated Perspective* 335–364 (2017), [https://doi.org/10.1007/978-3-319-59489-7\\_16](https://doi.org/10.1007/978-3-319-59489-7_16).
- [21] R.B. Chatfield, M. Sorek-Hamer, R.F. Esswein, A. Lyapustin, Satellite mapping of PM<sub>2.5</sub> episodes in the wintertime San Joaquin Valley: a “static” model using column water vapor, *Atmos. Chem. Phys.* 20 (2020) 4379–4397, <https://doi.org/10.5194/acp-2019-262>.
- [22] Y. Park, B. Kwon, J. Heo, X. Hu, Y. Liu, T. Moon, Estimating PM<sub>2.5</sub> concentration of the conterminous United States via interpretable convolutional neural networks, *Environ. Pollut.* 256 (2020) 113395, <https://doi.org/10.1016/j.envpol.2019.113395>.
- [23] H. Xia, et al., Micro-pulse upconversion Doppler lidar for wind and visibility detection in the atmospheric boundary layer, *Opt. Lett.* 41 (2016) 5218–5221, <https://doi.org/10.1364/ol.41.005218>.
- [24] T. Wei, H. Xia, B. Yue, Y. Wu, Q. Liu, Remote sensing of raindrop size distribution using the coherent Doppler lidar, *Opt. Express* 29 (2021) 17246–17257, <https://doi.org/10.1364/OE.426326>.
- [25] S. Lolli, Machine learning techniques for vertical lidar-based detection, characterization, and classification of aerosols and clouds: a comprehensive survey, *Remote Sens. (Basel)* 15 (2023) 4318, <https://doi.org/10.3390/rs15174318>.
- [26] S. Lolli, Urban PM<sub>2.5</sub> concentration monitoring: a review of recent advances in ground-based, satellite, model, and machine learning integration, *Urban Clim.* 63 (2025) 102566, <https://doi.org/10.1016/j.uclim.2025.102566>.
- [27] L. Lv, et al., Shipborne lidar measurements of ozone over the southeastern coastal regions of China in winter, *Environ. Res.* 121165 (2025), <https://doi.org/10.1016/j.envres.2025.121165>.
- [28] X. Shang, H. Xia, X. Dou, M. Shangguang, M. Li, C. Wang, Adaptive inversion algorithm for 1.5μm visibility lidar incorporating in situ Angstrom wavelength exponent, *Opt. Commun.* 418 (2018) 129–134, <https://doi.org/10.1016/j.optcom.2018.03.009>.
- [29] J. Xian, et al., Novel Lidar algorithm for horizontal visibility measurement and sea fog monitoring, *Opt. Express* 26 (2018) 34853, <https://doi.org/10.1364/oe.26.034853>.
- [30] H. Zhang, et al., Fitting of hygroscopic factor between PM<sub>2.5</sub> mass concentration and aerosol backscattering coefficient in Hefei area, *Chin. J. Lasers* 45 (2018) 163–169, <https://doi.org/10.3788/cj201845.0704006>, in Chinese.
- [31] B.-Y. Kim, J.W. Cha, Y.H. Lee, Estimation of PM<sub>10</sub> and PM<sub>2.5</sub> using backscatter coefficient of ceilometer and machine learning, *Aerosol Air Quality Res.* 23 (2023) 230033, <https://doi.org/10.4209/aaqr.230033>.
- [32] N. Meskhidze, et al., Improving estimates of PM<sub>2.5</sub> concentration and chemical composition by application of High Spectral Resolution Lidar (HSRL) and Creating Aerosol Types from chemistry (CATCH) algorithm, *Atmos. Environ.* 250 (2021) 118250, <https://doi.org/10.1016/j.atmosenv.2021.118250>.
- [33] A. Kolgotin, D. Müller, A. Romanov, Particle microphysical parameters and the complex refractive index from 3β+2α HSRL/Raman lidar measurements: conditions of accurate retrieval, retrieval uncertainties and constraints to suppress the uncertainties, *Atmos.* 14 (2023) 1159, <https://doi.org/10.3390/atmos14071159>.
- [34] H. Shao, et al., Exploring the Conversion Model from Aerosol Extinction Coefficient to PM<sub>1</sub>, PM<sub>2.5</sub> and PM<sub>10</sub> Concentrations, *Remote Sens. (Basel)* 15 (2023) 2742, <https://doi.org/10.3390/rs15112742>.
- [35] M. Zhen, et al., Application of a fusion model based on machine learning in visibility prediction, *Remote Sens. (Basel)* 15 (2023) 1450, <https://doi.org/10.3390/rs15051450>.
- [36] S. Kameyama, T. Ando, K. Asaka, Y. Hirano, S. Wadaka, Compact all-fiber pulsed coherent Doppler lidar system for wind sensing, *Appl. Opt.* 46 (2007) 1953–1962, <https://doi.org/10.1364/AO.46.001953>.
- [37] E. Päschke, R. Leinweber, V. Lehmann, An assessment of the performance of a 1.5 μm Doppler lidar for operational vertical wind profiling based on a 1-year trial, *Atmos. Meas. Tech.* 8 (2015) 2251–2266, <https://doi.org/10.5194/amt-8-2251-2015>.
- [38] T. Wei, M. Wang, K. Wu, J. Yuan, H. Xia, S. Lolli, Characterizing urban planetary boundary layer dynamics using 3-year doppler wind lidar measurements in a western Yangtze River delta city, China, *Atmos. Meas. Tech.* 18 (2025) 1841–1857, <https://doi.org/10.5194/amt-18-1841-2025>.
- [39] K. Kalmankoski, X. Shang, M. Komppula, J. Toivonen, Calibration of backscattering coefficients with coherent heterodyne lidar utilizing molecular scattering, *Opt. Express* 33 (2025) 3325–3338, <https://doi.org/10.1364/oe.543936>.
- [40] T. Wei, et al., Retrieving aerosol backscatter coefficient using coherent Doppler wind lidar, *Opt. Express* 33 (2025) 6832–6849, <https://doi.org/10.1364/oe.551730>.
- [41] M. Queißer, M. Harris, S. Knoop, Atmospheric visibility inferred from continuous-wave Doppler wind lidar, *Atmos. Meas. Tech.* 15 (2022) 5527–5544, <https://doi.org/10.5194/amt-15-5527-2022>.
- [42] T. Wei, et al., Simultaneous wind and rainfall detection by power spectrum analysis using a VAD scanning coherent doppler lidar, *Opt. Express* 27 (2019) 31235–31245, <https://doi.org/10.1364/OE.27.031235>.
- [43] T. Wei, H. Xia, Y. Wu, J. Yuan, C. Wang, X. Dou, Inversion probability enhancement of all-fiber CDWL by noise modeling and robust fitting, *Opt. Express* 28 (2020) 29662–29675, <https://doi.org/10.1364/OE.401054>.
- [44] M. Wang, et al., A long-term Doppler wind LIDAR study of heavy pollution episodes in western Yangtze River delta region, China, *Atmos. Res.* 310 (2024) 107616, <https://doi.org/10.1016/j.atmosres.2024.107616>.
- [45] E.Y. Boateng, J. Otoo, D.A. Abaye, Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review, *J. Data Anal. Inf. Process.* 8 (2020) 341–357, <https://doi.org/10.4236/jdaip.2020.84020>.
- [46] M. Filonchyk, M. Peterson, Development, progression, and impact on urban air quality of the dust storm in Asia in March 15–18, 2021, *Urban Clim.* 41 (2022) 101080, <https://doi.org/10.1016/j.uclim.2021.101080>.
- [47] C.E. Shi, X.L. Deng, Y.J. Yang, B.W. Wu, Y.J. Meng, Analyses on the causes of the persistent haze in Anhui Province in January 2013, *Climatic Environ. Res.* 19 (2014) 227–236, <https://doi.org/10.3878/j.issn.1006-9585.2014.13112>, in Chinese.
- [48] M.G. Elfeky, W.G. Aref, A.K. Elmagarmid, Periodicity detection in time series databases, *IEEE Trans. Knowledge Data Eng.* 17 (2005) 875–887, <https://doi.org/10.1109/tkde.2005.114>.

- [49] W. Claus, *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, Springer. ed., Springer, New York, NY, 2006.
- [50] A. Hagar, Extinction and visibility measurements in the lower atmosphere with UV-YAG-lidar. NASA. Langley Research Center 13th International Laser Radar Conference, 1986.
- [51] R. Vishnu, Y.B. Kumar, E.J.J. Samuel, Measurements of long range transport using two wavelength and polarization lidar over tropical rural site Gadanki (13.450 N, 79.170 E), *Lidar Remote Sens. Environ. Monitor.* XV 9879 (2016) 176–186, <https://doi.org/10.1117/12.2223686>.
- [52] H. Xia, et al., Long-range micro-pulse aerosol lidar at 1.5  $\mu\text{m}$  with an upconversion single-photon detector, *Opt. Lett.* 40 (2015) 1579–1582, <https://doi.org/10.1364/OL.40.001579>.
- [53] P. Zieger, R. Fierz-Schmidhauser, E. Weingartner, U. Baltensperger, Effects of relative humidity on aerosol light scattering: results from different European sites, *Atmos. Chem. Phys.* 13 (2013) 10609–10631, <https://doi.org/10.5194/acp-13-10609-2013>.
- [54] G. Titos, et al., Effect of hygroscopic growth on the aerosol light-scattering coefficient: a review of measurements, techniques and error sources, *Atmos. Environ.* 141 (2016) 494–507, <https://doi.org/10.1016/j.atmosenv.2016.07.021>.
- [55] G. Hänel, An attempt to interpret the humidity dependencies of the aerosol extinction and scattering coefficients, *Atmos. Environ.* 15 (1981) 403–406, [https://doi.org/10.1016/0004-6981\(81\)90045-7](https://doi.org/10.1016/0004-6981(81)90045-7).
- [56] R.A. Kotchenruther, P.V. Hobbs, D.A. Hegg, Humidification factors for atmospheric aerosols off the mid-Atlantic coast of the United States, *J. Geophys. Res. Atmos.* 104 (1999) 2239–2251, <https://doi.org/10.1029/98jd01751>.
- [57] J. Chen, et al., A parameterization of low visibilities for hazy days in the North China Plain, *Atmos. Chem. Phys.* 12 (2012) 4935–4950, <https://doi.org/10.5194/acp-12-4935-2012>.
- [58] S. Li, E. Joseph, Q. Min, B. Yin, R. Sakai, M.K. Payne, Remote sensing of PM<sub>2.5</sub> during cloudy and nighttime periods using ceilometer backscatter, *Atmos. Meas. Tech.* 10 (2017) 2093–2104, <https://doi.org/10.5194/amt-10-2093-2017>.
- [59] Tin, Kam, The random subspace method for constructing decision forests, *IEEE Trans. Pattern Analysis Mach. Intell.* 20 (1998) 832, <https://doi.org/10.1109/34.709601>.
- [60] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: *Preprints, Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [61] Ke, G., et al., 2017: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- [62] Y. Wu, S. Qi, F. Hu, S. Ma, W. Mao, W. Li, Recognizing activities of the elderly using wearable sensors: a comparison of ensemble algorithms based on boosting, *Sens. Rev.* 39 (2019) 743–751, <https://doi.org/10.1108/sr-11-2018-0309>.
- [63] J. Zhong, et al., Robust prediction of hourly PM<sub>2.5</sub> from meteorological data using LightGBM, *Natl. Sci. Rev.* 8 (2021) nwaa307, <https://doi.org/10.1093/nsr/nwaa307>.
- [64] Vapnik, V., 1998: The support vector method of function estimation. *Nonlinear modeling: Advanced black-box techniques*, Springer, 55–85.
- [65] Smola, A. J., B. Schölkopf, and computing, 2004: A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222, DOI: 10.1023/b:stco.0000035301.49549.88.
- [66] C.R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, L.J. Guibas, Volumetric and multi-view cnns for object classification on 3d data, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5648–5656, <https://doi.org/10.1109/cvpr.2016.609>.
- [67] P.Y. Kow, L.C. Chang, C.Y. Lin, C.C.K. Chou, F.-J. Chang, Deep neural networks for spatiotemporal PM<sub>2.5</sub> forecasts based on atmospheric chemical transport model output and monitoring data, *Environ. Pollut.* 306 (2022) 119348, <https://doi.org/10.1016/j.envpol.2022.119348>.
- [68] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, *Neurocomputing* 70 (2006) 489–501, <https://doi.org/10.1016/j.neucom.2005.12.126>.
- [69] X. Wang, R. Zhang, W. Yu, The effects of PM<sub>2.5</sub> concentrations and relative humidity on atmospheric visibility in Beijing, *J. Geophys. Res. Atmos.* 124 (2019) 2235–2259, <https://doi.org/10.1029/2018jd029269>.
- [70] P. Lara-Benitez, M. Carranza-García, J.C. Riquelme, An experimental review on deep learning architectures for time series forecasting, *Int. J. Neural Syst.* 31 (2021) 2130001, <https://doi.org/10.1142/s0129065721300011>.
- [71] P. Zieger, et al., Influence of water uptake on the aerosol particle light scattering coefficients of the central European aerosol, *Tellus B: Chem. Phys. Meteorol.* 66 (2014) 22716, <https://doi.org/10.3402/tellusb.v66.22716>.
- [72] W.S. Won, R. Oh, W. Lee, S. Ku, P.-C. Su, Y.-J. Yoon, Hygroscopic properties of particulate matter and effects of their interactions with weather on visibility, *Sci. Rep.* 11 (2021) 16401, <https://doi.org/10.1038/s41598-021-95834-6>.
- [73] X. Shang, H. Xia, X. Dou, M. Shangguan, M. Li, C. Wang, Adaptive inversion algorithm for 1.5  $\mu\text{m}$  visibility lidar incorporating in situ angstrom wavelength exponent, *Opt. Commun.* 418 (2018) 129–134, <https://doi.org/10.1016/j.optcom.2018.03.009>.
- [74] X. Luo, H. Bing, Z. Luo, Y. Wang, L. Jin, Impacts of atmospheric particulate matter pollution on environmental biogeochemistry of trace metals in soil-plant system: a review, *Environ. Pollut.* 255 (2019) 113138, <https://doi.org/10.1016/j.envpol.2019.113138>.
- [75] O.R. Omokungbe, et al., Analysis of the variability of airborne particulate matter with prevailing meteorological conditions across a semi-urban environment using a network of low-cost air quality sensors, *Heliyon* 6 (2020), <https://doi.org/10.1016/j.heliyon.2020.e04207>.
- [76] X. Zhao, X. Zhang, X. Xu, J. Xu, W. Meng, W. Pu, Seasonal and diurnal variations of ambient PM<sub>2.5</sub> concentration in urban and rural environments in Beijing, *Atmos. Environ.* 43 (2009) 2893–2900, <https://doi.org/10.1016/j.atmosenv.2009.03.009>.
- [77] R. Huang, L. Chun, Seasonal variation characteristics and forecasting model of PM<sub>2.5</sub> in Changsha, Central City in China, *J. Environ. Anal. Toxicol.* 7 (2017) 429–435, <https://doi.org/10.4172/2161-0525.1000429>.
- [78] E. Swietlicki, et al., Hygroscopic properties of submicrometer atmospheric aerosol particles measured with H-TDMA instruments in various environments—a review, *Tellus B: Chem. Phys. Meteorol.* 60 (2008) 432–469, <https://doi.org/10.3402/tellusb.v60i3.16936>.
- [79] M. Tang, et al., A review of experimental techniques for aerosol hygroscopicity studies, *Atmos. Chem. Phys.* 19 (2019) 12631–12686, <https://doi.org/10.5194/acp-19-12631-2019>.
- [80] C. Wang, et al., Relationship analysis of PM<sub>2.5</sub> and boundary layer height using an aerosol and turbulence detection lidar, *Atmos. Meas. Tech.* 12 (2019) 3303–3315, <https://doi.org/10.5194/amt-12-3303-2019>.
- [81] X.Y. Zhang, Y. Wang, T. Niu, X. Zhang, S. Gong, Y. Zhang, J. Sun, Atmospheric aerosol compositions in China: spatial/temporal variability, chemical signature, regional haze distribution and comparisons with global aerosols, *Atmos. Chem. Phys.* 12 (2012) 779–799, <https://doi.org/10.5194/acp-12-779-2012>.
- [82] L. Zhang, S. Gong, J. Padro, L. Barrie, A size-segregated particle dry deposition scheme for an atmospheric aerosol module, *Atmos. Environ.* 35 (2001) 549–560, [https://doi.org/10.1016/S1352-2310\(00\)00326-5](https://doi.org/10.1016/S1352-2310(00)00326-5).
- [83] B. Chen, An interpretable physics-informed deep learning model for estimating multiple air pollutants, *Giscience Remote Sens.* 62 (2025) 2482272, <https://doi.org/10.1080/15481603.2025.2482272>.