https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


# Inversion Algorithm of Black Carbon Mixing State Based on Machine Learning

Zeyuan Tian[1,2], Jiandong Wang[1,2], Jiaping Wang[3,4], Chao Liu[1,2], Jinbo Wang[3,4], Zhouyang Zhang[1,2], Yuzhi Jin[1,2], Sunan Shen[1,2], Bin Wang[1,2], Wei Nie[3,4], Xin Huang[3,4], Aijun Ding[3,4]

[1]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing, China,
[2]China Meteorological Administration Aerosol-Cloud-Precipitation Key Laboratory, School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing, China,
[3]Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing, China,
[4]National Observation and Research Station for Atmospheric Processes and Environmental Change in Yangtze River Delta, Nanjing, China,

*Correspondence to*: Jiandong Wang (jiandong.wang@nuist.edu.cn)

**Abstract.** Black carbon (BC) radiative impact is significantly influenced by its mixing state. Single-particle soot photometer (SP2) is a widely recognized instrument for quantifying BC mixing state. However, the derivation of BC mixing state from SP2 is quite challenging. Since the SP2 records individual particle signals, it requires complex data processing to convert raw signals into particle size and mixing states. Besides, the rapid accumulation of substantial data volumes impedes real-time analysis of BC mixing states. This study employs a light gradient boosting machine (LightGBM) to establish an inversion model which directly correlates SP2 signals with the mixing state of BC-containing particles. Our model achieves high accuracy for both particle size inversion and optical cross-section inversion of BC-containing particles, with $R^2$ higher than 0.98. Further, we employed the SHapley Additive exPlanation (SHAP) method to analyze the importance of input features from SP2 signals in the inversion model of the entire particle diameter ($D_p$) and explored their underlying physical significance. Compared to the widely used Leading-Edge-Only (LEO) fitting method, the machine learning (ML) method utilizes a larger coverage of signals encompassing the peak of scattering signal rather than the leading-edge data. This allows for more accurate capture of the diverse characteristics of particles. Moreover, the ML method uses signals with a high signal-to-noise ratio, providing better noise resistance. Our model is capable of accurately and efficiently acquiring the single-particle information and statistical results of the BC mixing state, which provides essential data for BC aging mechanism investigation and further BC radiative effects assessment.

## 1 Introduction

Black carbon (BC) is the dominant absorbing aerosol, making it an important contributor to positive radiative forcing in the present-day atmosphere (Bond et al., 2013; Bond and Bergstrom, 2006; Fierce et al., 2020; Liu et al., 2017; Matsui et al., 2018; Ramanathan and Carmichael, 2008). As the product of incomplete combustion of fossil fuels combustion and biomass burning

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


(Jacobson, 2001), BC is refractory with a vaporization temperature near 4000 K. Because of the coagulation and condensation with other aerosol components during atmospheric transport, freshly emitted BC changes from externally mixed state to internally mixed structure. Changes in the mixing state can alter the light absorption and other properties of BC, thereby affecting its climate effect. For example, the presence of coating on BC can increase its mass absorption cross-section (MAC) relative to uncoated BC by lensing effect (Bond and Bergstrom, 2006; Cappa et al., 2012; Fuller et al., 1999). Therefore, identifying the mixing states of BC-containing particles and their relative abundance is essential for evaluating their climate effects.

The single-particle soot photometer (SP2) is a well-recognized instrument that can be used for measuring the mixing state of BC (Moteki and Kondo, 2007; Schwarz et al., 2006; Sedlacek et al., 2012; Stephens et al., 2003). By analyzing the signals observed by SP2, quantitative characterization of the mixing states of BC can be obtained. Because the non-refractory material vaporizes due to the absorption of laser energy by the BC core, the scattering signals of BC-containing particles obtained by SP2 will be distorted, which poses significant difficulties in analyzing the original particle size ($D_p$). The leading-edge-only (LEO) fitting method is widely used (Gao et al., 2007; Moteki and Kondo, 2008; Schwarz et al., 2008) to obtain $D_p$, wherein the complete Gaussian function is reconstructed by fitting the scattering signal before particle vaporization (Liu et al., 2014; Shiraiwa et al., 2008; Zhang et al., 2016). Since SP2 can track the incandescence and scattering signal of each particle, field observation using SP2 will generate a large amount of data. Performing physical inversion of particle size requires complex data processing and fitting processes, making it difficult to obtain real-time online BC mixing states. As an alternative, data-driven models can provide a good supplement to physical process-based models. Machine learning (ML) is a rapidly developing data-driven model which can efficiently simulate the nonlinear relationship between input and output, and is widely used in various fields (Carleo et al., 2019; Jordan and Mitchell, 2015; Liakos et al., 2018; Tarca et al., 2007). Applying ML to the inversion of BC mixing states can efficiently process a large number of SP2 datasets.

In this study, an inversion model is built using the light gradient boosting machine (LightGBM) to associate the SP2 time-dependent signals with the size of individual BC-containing particles and their optical properties. This method can simplify the process of quantitative analysis of BC mixing states, making it capable of real-time mixing state analysis. In addition, the SHapley Additive explanation (SHAP) approach is introduced to quantify the individual effect of signal factors on prediction. The BC mixing state inversion model developed in this study is also compared with the LEO fitting method. Finally, based on our inversion model, the mixing state characteristics of BC-containing particles can be analyzed in detail, including both single-particle scale and statistical features.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


## 2 Method

### 2.1 Experimental site

The SP2 observational data used in this study is from 1 April 2022 to 31 May 2022 at SORPES (Station for Observing Regional Process of the Earth system) station, which located in Xianlin Campus of Nanjing University in Nanjing, Jiangsu Province (a regional background site in the Yangtze River Delta region in China).

### 2.2 SP2 apparatus and detection principle

The SP2 consists of an intracavity Nd:YAG laser and four optical detectors. The laser operates in a TEM00 mode, with a Gaussian intensity distribution. The laser intensity within the cavity is approximately $10^6$ W $cm^{-2}$, which is sufficient to vaporize absorbing particles as they pass through the beam (Stephens et al., 2003). The refractory particle absorbs light and has a high vaporization temperature. When heated in the laser beam to the boiling point (about 4000 K), it emits visible thermal radiation (''incandescent light''). The intensity of this thermal radiation depends on the composition and quality of the refractory components, regardless of the particle morphology and mixing state (Schwarz et al., 2006; Slowik et al., 2007). Pure scattering particles cannot absorb enough energy to heat themselves and therefore do not emit incandescent light. They are sized based on the amount of light they scatter from the laser, which exhibits a Gaussian dependence with time.

Four optical detectors are synchronously sampled at 5 MHz. One avalanche photo-detector (APD) is optically filtered to pass only 1064 nm radiation and measures the scattering signal from all particles, including both pure scattering particles and absorbing particles. The two other APDs measure incandescence signal in the visible range, optically filtered to pass broadband light at 400–650 nm and narrowband light at 610–650 nm. The ratio of signals from these two detectors can be employed to ascertain the vaporization temperature of the particles (Schwarz et al., 2006), ensuring that the measured particle is BC. The fourth two-element APD (TEAPD) detector measures the location of leading-edge data in the laser beam, which can be used to analyze the amount of coating or mixing state of the incandescent particles.

## 3 Machine-learning-based inversion algorithm

### 3.1 Construction of feature dataset

Supervised machine learning is a widely employed technique wherein algorithms are utilized to derive a function that maps inputs to desired outputs. This study aims to establish the relationship between SP2 signals and particle size as well as optical properties, where particle size includes the entire particle diameter ($D_p$) and BC core diameter ($D_c$), and optical properties encompass scattering cross-section ($C_{sca}$) and absorption cross-section ($C_{abs}$). The supervised ML is utilized in this study to achieve this goal, as depicted in Fig. 1. The learning process within an ML model typically comprises two steps: training and testing. In the training process, it is necessary to construct pre-processed datasets, including inputs (feature dataset) and outputs

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


(label dataset). In this study, the SP2 signals are used as the input data for the ML model, incorporating scattering signals and incandescence signals. Each of these signals is represented as 100-dimensional data, containing information that determines particle size and optical properties. However, directly using such high-dimensional data for ML would result in poor model performance. Additionally, the original signals also contain some instrument background signals, which can interfere with the learning process and affect the accuracy and efficiency of the model.

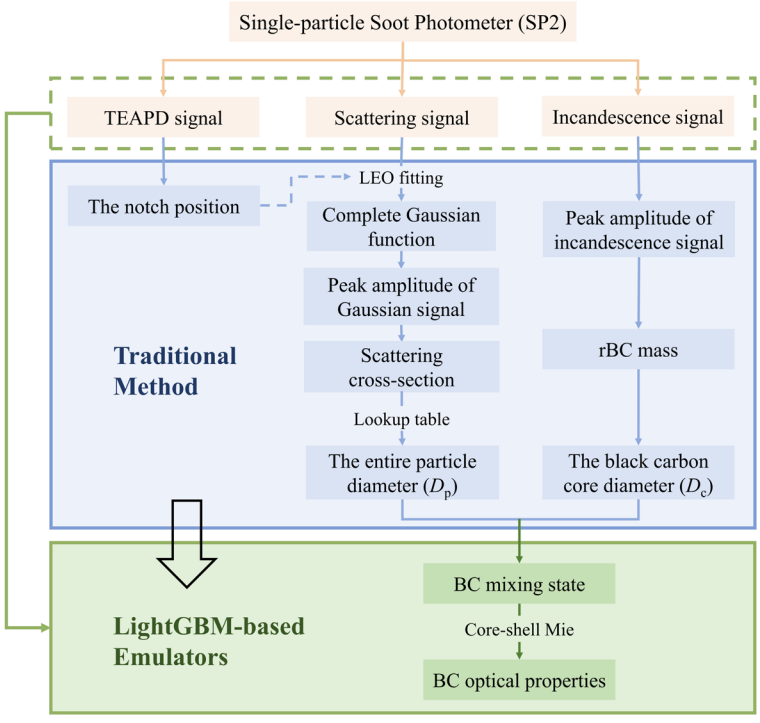


**Figure 1.** Schematic diagram of BC mixing state inversion process.

To enhance the ML performance, it is imperative to preprocess the original signals to reduce data dimension and eliminate unnecessary noise. This usually entails procedures of feature selection and feature extraction, aiming to identify and retain the most relevant signal features related to particle size and optical properties. Since different types of particles have different effective SP2 signal dimensions, it is crucial to pre-classify the particle types and select the appropriate signals for each type of particle as the feature data input for the ML model.

In this study, ambient particles are classified into pure scattering particles and BC-containing particles (both externally mixed BC and internally mixed BC). Firstly, we differentiate the pure scattering particle and BC-containing particle depending on whether it has the incandescence signal. When the peak height of the incandescence signal exceeds a certain degree, the particle will be considered as BC-containing particle. Otherwise, it will be considered as pure scattering particle (Fig. 2a). According to the time delay ($\Delta t$), namely the time difference of the peak of the incandescence signal and the scattering signal (Moteki

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


and Kondo, 2007), BC-containing particles are further classified into externally mixed BC ($\Delta t < 1$ μs) (Fig. 2b) and internally mixed BC ($\Delta t \geq 1$ μs) (Fig. 2c).

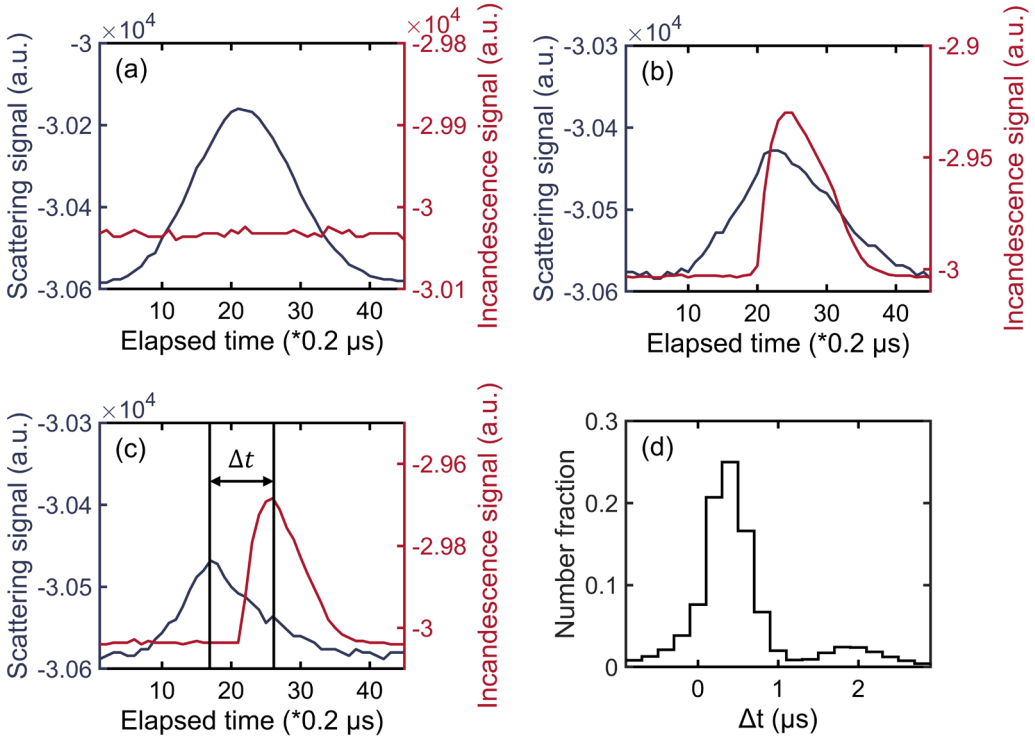


**Figure 2.** (a-c) Time series of the scattering signal and incandescence signal of different types of particles: (a) pure scattering particle; (b) externally mixed BC; (c) internally mixed BC. (d) Histogram of the time delay ($\Delta t$) of BC-containing particles.

The inversion of pure scattering particles involves the $D_p$ and $C_{sca}$. Both of these parameters are calculated through the physical inversion process by using the scattering signals obtained by SP2. Therefore, when constructing an ML model for pure scattering particles, only the scattering signals are used as feature data. The scattering signals of particles when they pass through the laser beam are influenced by two factors: the particle size and their position in the SP2. To facilitate the comparison of signal intensities produced by different particles, it is essential to ensure that these particles are located in the same position. Due to the constant sample flow, the zero-crossing point in the TEAPD signal can indicate the position of particles in the instrument. That is, at the zero-crossing point, particles are in the same position in the instrument. Since the SP2 records the data from all four detector channels simultaneously, this reference position is valid for the other three detector signals. For pure scattering particles, the 45-dimensional scattering signals near the zero-crossing point for each particle are selected as the feature data.

To perform the $D_c$ inversion of BC-containing particles, the incandescence signals are employed as the input feature for the model since $D_c$ is directly related to the incandescence signals of the particles. Differing from the feature data selection method

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


in the inversion model for pure scattering particles, the incandescence signal is not influenced by the particle's position in the SP2. Therefore, the 45-dimensional incandescence signals near the peak of each particle's incandescence signal are used as the feature data for model establishment. Similarly, the incandescence signals of particles are also adopted in the model to invert the optical properties of externally mixed BC, constructing feature data in the same way.

As for internally mixed BC, when it passes through the laser beam, it both absorbs and scatters laser light. During the process of physical inversion, it is essential to utilize the scattering and incandescence signals to obtain $D_p$ and optical properties of internally mixed BC. Therefore, when constructing an ML inversion model for internally mixed BC, both the scattering and incandescence signals need to be considered. This study employs the 90-dimensional feature data for internally mixed BC, consisting of 45 dimensions near the zero-crossing point for both the scattering and incandescence signals.

### 3.2 Construction of label dataset

SP2 data have been used to optically size particles. This study uses particle size and optical properties obtained from traditional physical inversion methods as the label dataset for the ML model. For pure scattering particles, when they pass through the laser, the particle size does not change, resulting in an undistorted scattering signal. Therefore, their $D_p$ is positively correlated with the scattering amplitude. The Mie calculations indicate that, for spherical particles with diameters less than 1 μm, the scattering amplitude detected by SP2 exhibits a monotonic relationship with scattering cross-section (Gao et al., 2007). The constant ratio between them is determined by the calibration using polystyrene latex spheres (PSL). The $D_p$ of pure scattering particles can be obtained by correlating the scattering cross-section of particles with known particle size.

When the BC-containing particle enters SP2, BC will absorb the energy of the laser and emit significant incandescent light. The peak intensity of thermal radiation from a particle is proportional to its refractory BC mass ($M_{BC}$) (Moteki and Kondo, 2007). According to the empirical relationship between the incandescent light intensity and the particle mass calibrated using fullerene soot, the $M_{BC}$ of each BC-containing particle can be quantified by the peak height of the incandescence signal. Subsequently, the measured $M_{BC}$ can be further converted into the mass-equivalent diameter $D_c$ assuming a density of 1.8 g $cm^{-3}$ (Bond and Bergstrom, 2006).

Because the BC component in the particles absorbs the energy of the laser, the coating will vaporize after being heated to boiling temperature. The particle size decreases gradually due to vaporization, and the scattering signal deviates from the Gaussian function. The leading-edge-only (LEO) fitting method is used to reconstruct the Gaussian signal. The $D_p$ of a BC-containing particle can be derived by inputting the LEO fitted scattering signal and $D_c$ into Mie calculations.

Based on the particle size calculations, the optical properties of particles can be further obtained. For pure scattering particles and externally mixed BC, Mie scattering theory can be used to calculate the $C_{sca}$ and $C_{abs}$ with known refractive index and

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


optical size of particles. For internally mixed BC, a core-shell model is required, that is, they are considered to have an ideal BC core and a uniform non-absorbing coating material. The Mie scattering algorithm of core-shell structured particles can be used to obtain the optical cross-section of internally mixed BC. In the above calculation process, a complex refractive index of 1.95 + 0.96i for BC core (Moteki et al., 2023), and a complex refractive index of 1.5 + 0i (Schnaiter et al., 2005) for the coating of the internally mixed BC and pure scattering particles are used.

Considering the low signal-to-noise ratio for small particles in SP2, this study set the lower limit for the $D_p$ of pure scattering particles at 170 nm. For BC-containing particles, the lower limits for $D_c$ and $D_p$ are set as 90 nm and 120 nm, respectively. The upper limit for all particle sizes is set as 600 nm. When the particle is too large, the resulting signal exceeds the SP2 detection threshold, leading to incomplete signal recording. Therefore, even though the original particle sizes can be obtained through the LEO fitting method, these particles are not included in the ML dataset. This preprocessing step ensures the quality of the data for ML, thereby improving the accuracy of the model's prediction.

### 3.3 Machine learning model

To quickly inverse the particle size and optical properties of particles detected by SP2, LightGBM (Ke et al., 2017) is used in this study. LightGBM is a novel GBDT (Gradient Boosting Decision Tree) algorithm. In resemblance to GBDT, the objective output of each tree is determined by the discrepancy between the prediction of the tree model and the expected output from the preceding tree, while the input remains unchanged. A collection of trees is employed to make predictions, resulting in the final prediction. Different from traditional GBDT algorithms, LightGBM uses a histogram-based algorithm to avoid calculating all continuous features and takes discrete bins as the unit, which consumes less memory and reduces the complexity of data separation to speed up the training process (Fan et al., 2019). In addition to the histogram algorithm, LightGBM adapts the leaf-wise strategy to grow trees, identifying the leaf with the maximum gain in split variance to perform the split, which is greedier than the level-wise strategy (Gan et al., 2021). Furthermore, LightGBM incorporates gradient-based one-sided sampling and exclusive feature bundling (Sun et al., 2020), addressing large volumes of data instances and numerous features, respectively.

In the LightGBM model, many hyperparameters are used, which can be adjusted to improve the performance of the model in different applications. Table 1 lists the hyperparameters adjusted in this study and their related meanings. In this study, the SP2 data from May 11, 2022, to May 25, 2022, is used as the dataset for establishing the model. The data set is divided into a training set and a testing set (7:3). The training set is used to train the LightGBM regression model, while the testing set is used to evaluate the accuracy of the model. Based on the specified hyperparameter range, use the GridSearchCV function to form all possible parameter combinations and then perform a 5-fold cross-validation process to search for the optimal hyperparameter combination. Additionally, the early stop mechanism is added to prevent the model from overfitting. All the

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


optimized hyperparameters are listed in Table 2. With the hyperparameters in Table 2, the final LightGBM model can be trained.

**Table 1.** The main hyperparameters of the LightGBM model tuned in this study.

| Hyperparameters | Description |
|---|---|
| learning_rate | Control the shrinkage rate. |
| num_leaves | Control the maximum number of leaves of a decision tree. |
| max_bin | Control the max number of bins (data intervals) when the dataset of a parameter in the input layer is transformed to a histogram. |
| max_depth | Limit the max depth for a tree model. |
| feature_fraction | The proportion of the selected parameters to the total number of the parameters in the input layer. |
| bagging_fraction | The proportion of the selected data to the total data size. |
| bagging_freq | The frequency of re-sampling the data when bagging_fraction is smaller than 1.0. |

**Table 2.** The optimal hyperparameters for each particle type. The content in parentheses following the particle type name indicates the physical quantity that needs to be inverted for that type of particle.

| Hyperparameters | Particle Type | | | |
|---|---|---|---|---|
| | Pure scattering particle ($D_p$ / $C_{sca}$) | Externally mixed BC ($C_{sca}$ / $C_{abs}$) | Internally mixed BC ($D_p$ / $C_{sca}$ / $C_{abs}$) | BC-containing particle ($D_c$) |
| learning_rate | 0.05 | 0.05 | 0.05 | 0.05 |
| num_leaves | 30 | 50 | 700 | 45 |
| max_bin | 800 | 800 | 500 | 800 |
| max_depth | 15 | 15 | 50 | 15 |
| feature_fraction | 0.9 | 0.9 | 0.7 | 0.9 |
| bagging_fraction | 0.9 | 0.7 | 0.8 | 0.9 |
| bagging_freq | 2 | 3 | 4 | 2 |

### 3.4 Model performance evaluation

To establish a comprehensive understanding of the BC mixing state inversion model, it is necessary to evaluate and interpret it. The evaluation metrics employed in the inversion results of the BC mixing states in this paper include the coefficient of $R^2$, root mean square error (RMSE), and mean absolute error (MAE). $R^2$ is the most important index to verify the accuracy of the predicted result of a regression algorithm, with a range of 0 to 1. The result of the $R^2$ value equalling to 1 represents the regression model gives predictions without any error. In general, the higher the $R^2$ is, the better the fitting result is. RMSE stands for the standard deviation of the residuals between the predicted value and actual value calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\left(\sum_{i=1}^{m}(y_i - \hat{y}_i)^2\right)}, \quad (1)$$

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


where $m$ is the number of samples, $y_i$ is the actual value and $\hat{y}_i$ is the predicted value of $i^{th}$ sample. MAE is another statistical measure to evaluate the bias between predicted value and actual value, which is defined as:

$$\text{MAE} = \frac{1}{m}|y_i - \hat{y}_i| \,. \quad (2)$$

In general, the lower RMSE and MAE values represent the better fitting results of the model.

### 3.5 Model explanation

Given the inherent "black box" nature of machine learning models, comprehending the impact of input parameters on prediction results becomes challenging. As the model complexity increases, the need for post hoc explanations arises. Therefore, we involved the SHAP method to reveal the underlying reasoning behind predictions.

SHAP is a novel model interpretation method that uses the Shapely value from game theory to combine optimal credit allocation with local explanations (Lundberg and Lee, 2017). It can be used in conjunction with different ML models for model interpretation. Tree-SHAP (Lundberg et al., 2019) is used in the present study to determine. It uses a linear explanatory model and Shapley values to estimate the initial prediction model, as defined by Eq. (3):

$$f(x) = \Phi_0 + \sum_{i=1}^{p} \Phi_i \,, \quad (3)$$

where $f(x)$ represents the machine learning model's prediction; $\Phi_0$ is the base value of the model, which denotes the average prediction of all inputs; $\Phi_i$ is the SHAP value for feature $i$, indicating the contribution of feature $i$ to the prediction; and $p$ is the total number of features. The SHAP values provide a unified measure of feature importance, allowing for a detailed understanding of the impact of each feature on the model's output.

## 4 Result

### 4.1 Inversion results of particle size

The testing data are used to examine the model's accuracy. Figure 3 shows the particle size inversion results for different types of particles. Overall, the LightGBM model can successfully reproduce the $D_c$ and $D_p$ of particles. The $D_c$ inversion of BC-containing particles shows the best performance (Fig. 3b), with the $R^2$ value reaching 0.99. Additionally, the RMSE and MAE values are 0.36 nm and 0.14 nm, respectively, which are the smallest among the three particle size inversion models, as the incandescence signal peak height of BC-containing particles is linearly correlated with the mass of refractory BC. The $D_p$ inversion results for pure scattering particles are also satisfactory (Fig. 3a), with the $R^2$ of 0.99. The RMSE and MAE values are 1.2 nm and 0.64 nm, slightly lower than the $D_c$ inversion for BC-containing particles. The $D_p$ of pure scattering particles is calculated from the peak height of their scattering signals. Fluctuations in the instrument's voltage can lead to changes in laser intensity, which in turn affects the peak height of the scattering signals. This can result in slight deviations between the model-predicted values and actual values for certain particles. A few particles with prediction values that significantly deviate

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


from the actual values (> 30 nm) are affected by instrument noise signals, leading to abnormal scattering signals for these particles.

The $R^2$, RMSE, and MAE values for the $D_p$ inversion model of internally mixed BC are 0.98, 5.85 nm, and 2.98 nm, respectively (Fig. 3c). According to density distribution, the predicted values for the majority of particles are close to the actual values. Compared to the particle size inversions of the first two types, the relationship between $D_p$ and scattering signals and incandescence signals for internally mixed BC is nonlinear, making the physical inversion process more complex and involving more input variables, thus increasing the difficulty of inversion. The LEO fitting method and ML method adopts different parts of raw signals, leading to diverse $D_p$. The detailed discussion of this deviation is shown in Section 4.3. Furthermore, not all particle signals are in ideal conditions, and when two particles pass through the laser simultaneously, some interfering particle signals may be produced.

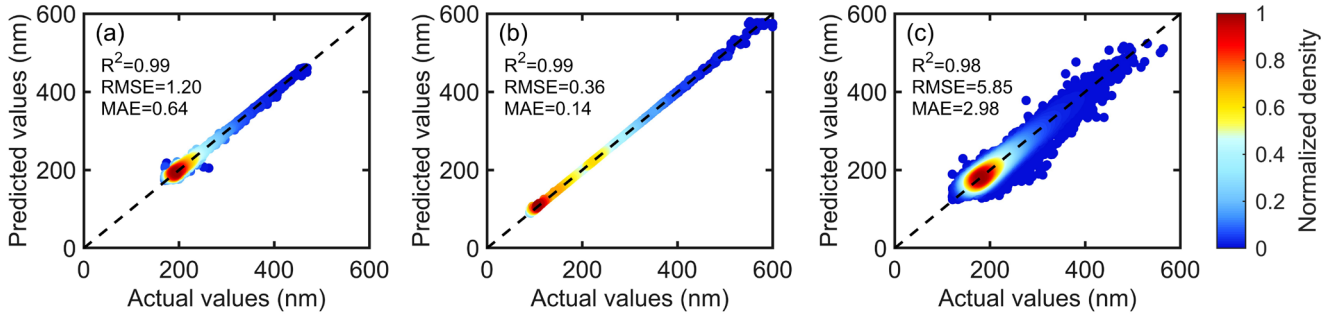


**Figure 3.** Inversion results of particle size for different types of particles: (a) $D_p$ of pure scattering particles; (b) $D_c$ of BC-containing particles; (c) $D_p$ of internally mixed BC.

### 4.2 Inversion results of optical properties

The results of the scattering and absorption cross-section inversion for three types of particles are shown in Figs. 4 and 5. Overall, the inversion of $C_{sca}$ and $C_{abs}$ of different types of particles shows a good performance, with $R^2$ higher than 0.98. For pure scattering particles, the $R^2$ of the $C_{sca}$ inversion is 0.99 (Fig. 4a), indicating close agreement between model predictions and actual values. For externally mixed BC, the $R^2$ for both the $C_{sca}$ and $C_{abs}$ inversion can reach 0.99 (Fig. 4b). The slight discrepancy between the model predictions and actual values for externally mixed BC with high optical cross-sections arises from the limited data sample, resulting in inadequate learning for these large particles. As for internally mixed BC (Fig. 4c), the $R^2$ values for the $C_{sca}$ and $C_{abs}$ inversion models are 0.98 and 0.99, respectively, with better inversion results for the $C_{abs}$. Except for a few particles with substantial deviations, the majority of particles exhibit high consistency between model predictions and actual value.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


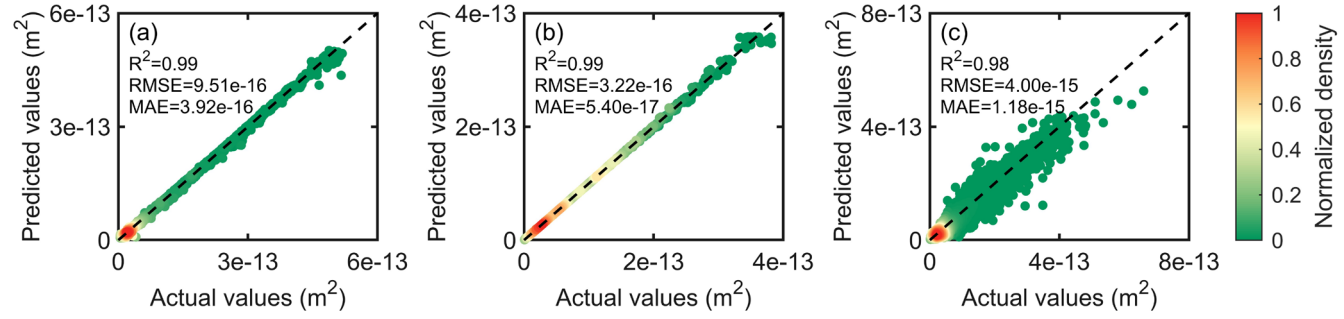


**Figure 4.** Inversion results of $C_{sca}$ for three types of particles：(a) pure scattering particles; (b) externally mixed BC; (c) internally mixed BC.

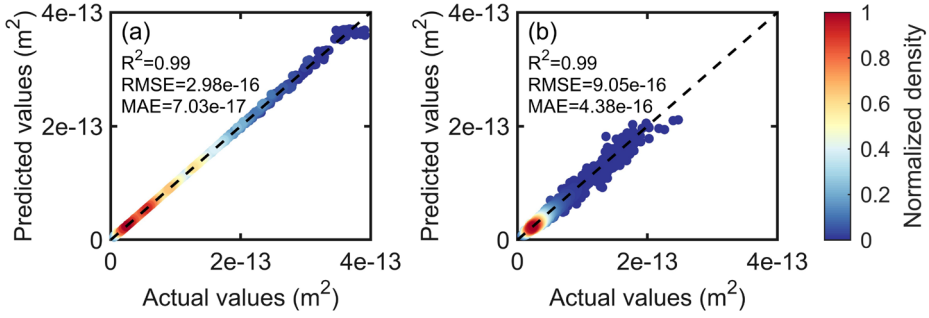


**Figure 5.** Inversion results of $C_{abs}$ for externally mixed BC (a) and internally mixed BC (b).

### 4.3 SHAP interpretations

To elucidate the relative contributions of various input features to the model output, the SHapley Additive exPlanation (SHAP) method was introduced to conduct feature importance analysis. Figure 6 presents the analysis results, where each point represents an individual data point from the dataset. The color of each point indicates the corresponding feature value for that specific sample, transitioning from blue to red as the feature value increases. The features are ranked by importance on the y-axis, with higher positions indicating greater importance. The position of an instance on the x-axis represents its SHAP value, which measures the impact of a feature on the model output for that specific data point. Points on the positive side of the zero line indicate a positive contribution of the input feature on the model prediction, whereas points on the negative side indicate a negative effect of the corresponding input feature. Additionally, the distance of a data point from the zero line is proportional to the impact of the corresponding input feature on the model prediction for that specific data point. To simplify the expression, the 45-dimensional scattering signals of the input model are designated as SCLG1, SCLG2, ..., SCLG45, and the 45-dimensional incandescent signals are named similarly as BBLG1, BBLG2, ..., BBLG45.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


Figure 6a shows the SHAP summary plot for the $D_p$ inversion model of internally mixed BC. The top fifteen features contributing to the model are listed. As can be seen, these features include eight scattering signal features and seven incandescence signal features. The specific distribution of these features within the signals is shown in Fig. 6b. According to the SHAP summary plot, the top three important features are SCLG22, SCLG23, and SCLG21, which correspond to three consecutive scattering signal positions. Similarly, the features SCLG15, SCLG14, and SCLG16 ranked 7th, 11th, and 12th respectively form a continuous scattering signal. These six features are located near the peak of the scattering signal, and as their feature values increase, the predicted $D_p$ also increases according to the change in SHAP values. In addition, among the top fifteen important features, the other two isolated scattering signal features are SCLG11 and SCLG1. The SHAP value for SCLG11 shows a similar trend to the aforementioned features as the feature value changes, while the SHAP value for SCLG1 can be either positive or negative as the feature value increases. The reason is that SCLG1 is at the starting position of the scattering signal input to the ML model, close to the baseline of the original scattering signal without dimensionality reduction, and is greatly interfered by instrument noise, thus having no clear correlation with the model's final output.

The seven incandescence signal features among the top fifteen important features form a continuous segment of the incandescence signal, namely BBLG24 to BBLG30. This segment corresponds to the position from the onset of incandescence signal emission by BC-containing particles to near the peak of the incandescence signal. The SHAP values for these seven incandescence signal features all exhibit a trend wherein their contribution to the prediction of $D_p$ transitions from positive to negative as the feature values increase. This is because for internally mixed BC with the same scattering cross-section, a larger $D_c$ leads to a thinner coating thickness, resulting in a smaller predicted $D_p$, given that the scattering coefficient of the BC core is greater than that of the coating. Furthermore, the peak height of the incandescence signal is positively correlated with $D_c$. Therefore, larger feature values correspond to smaller predicted $D_p$. Additionally, for internally mixed BC with the same $D_c$, the peak height of the incandescence signal remains the same, but the peak occurrence time varies with the thickness of the coating. When the coating is thicker, it takes longer for the coating to evaporate, resulting in a delayed onset of the incandescence signal. Compared to thinner coatings, this causes more data points from BBLG24 to BBLG30 to lie near the baseline of the incandescence signal, leading to smaller corresponding values (Fig. S1).

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


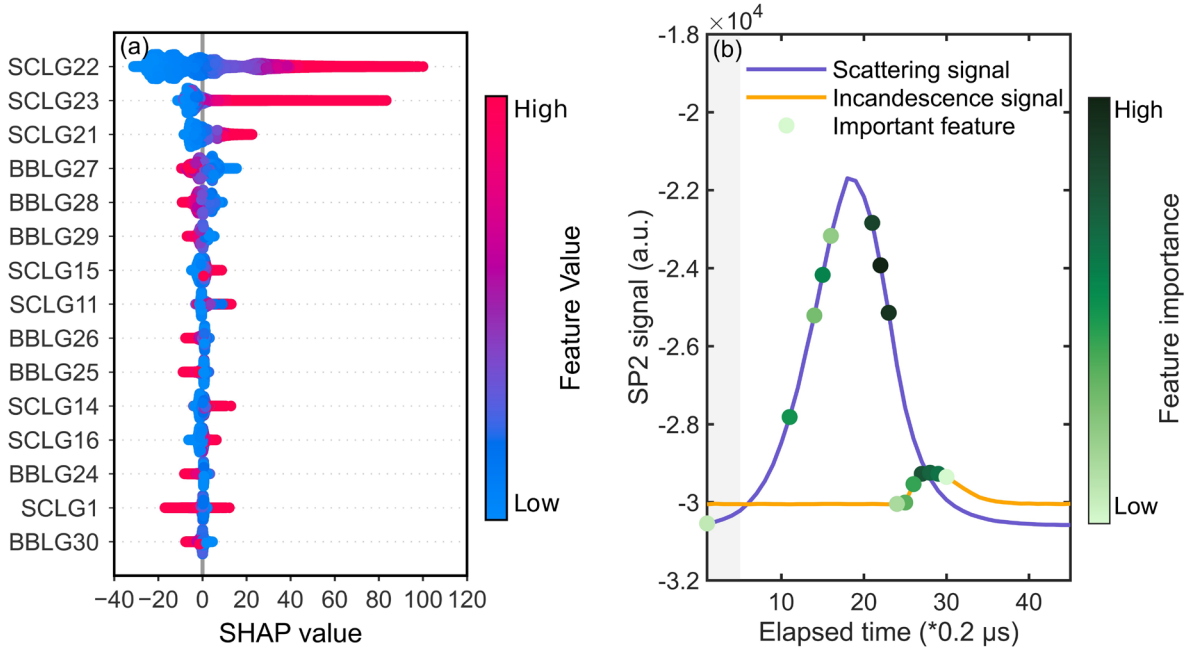


**Figure 6.** (a) The SHAP summary plot for the $D_p$ inversion model of internally mixed BC, showing the top fifteen features ranked by importance. (b) The specific positions of the top fifteen important features indicated by SHAP values within the scattering and incandescence signals input to the ML model. The darker color of the scatter points represents the higher importance ranking of the corresponding features. The gray shaded area shows the portion of the scattering signal used for LEO fitting that is included in the input features for ML.

According to the contribution of each feature indicated by SHAP values, it can be observed that the important features in the ML model differ from the leading-edge data used during the physical inversion process. ML model uses the signal near the peak, as illustrated in Fig. 6b, while the LEO fitting method uses the signal as the BC-containing particle enters the edge of the laser, prior to coating evaporation, to derive the particle size. Figure 7 displays the LEO fitting results for two distinct BC-containing particles. Since the data used for LEO fitting are almost identical, the fitting results in two Gaussian functions with the same distribution, yielding the same derived $D_p$. However, looking into the whole scattering signal, these two BC-containing particles differ significantly. The scattering signal adopted by the ML model can reflect this difference, resulting in different $D_p$ outcomes. Moreover, the leading edge is defined as the data from zero to 5 % of the maximum laser intensity in practice, with the baseline subtracted (Taylor et al., 2015). As shown in Fig. 7, this portion of the signal (in the grey-shaded area) is close to the baseline, making it more susceptible to noise interference. Compared to LEO fitting method, the ML model utilized a larger coverage of signals with high signal-to-noise ratio, providing better noise resistance.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


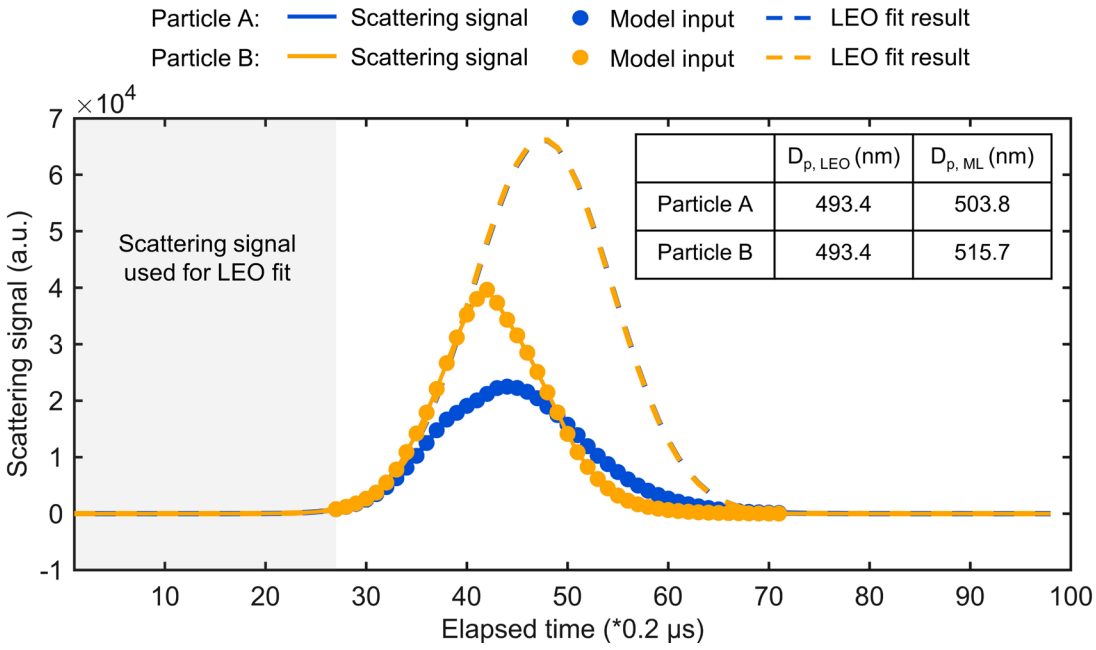


**Figure 7.** The comparison of the scattering signal used in the $D_p$ inversion process for internally mixed BC and specific calculation results between the LEO fitting method and the ML method. The solid line represents the scattering signal obtained by SP2, the part marked with solid dots is the scattering signal input to the ML model, the gray shaded area shows the leading-edge data used in the LEO fitting process, and the dashed line represents the scattering signal of the original particle reconstructed by LEO fitting.

### 4.4 Model application

The BC mixing state inversion model developed in this study exhibits broad applicability and can be applied to SP2 data obtained over different observation periods. To validate the model's effectiveness, we applied it to the SP2 dataset from April 2022. The results indicate that the model can rapidly and accurately invert the single-particle size information of BC-containing particles. Specifically, the model achieved an $R^2$ value of 0.99 for $D_c$ inversion and 0.98 for $D_p$ inversion (Table S1). Based on the inversion data, we can analyze the overall size distribution of internally mixed BC in April. From the number size distributions of the Fig. 8, it is evident that $D_c$ is primarily concentrated around 130 nm, while $D_p$ is mainly around 185 nm. The specific size distribution of particles in the two-dimensional histogram indicates that when $D_c$ is relatively small (<100 nm), $D_p$ is primarily distributed around 180 nm. When $D_c$ is small, some particles with thin coatings are not detected due to the detection limits of the SP2, resulting in overall thicker coating. As $D_c$ gradually increases to the range of 120 to 140 nm, the main distribution area of $D_p$ decreases to around 165 nm, corresponding to a reduction in coating thickness.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


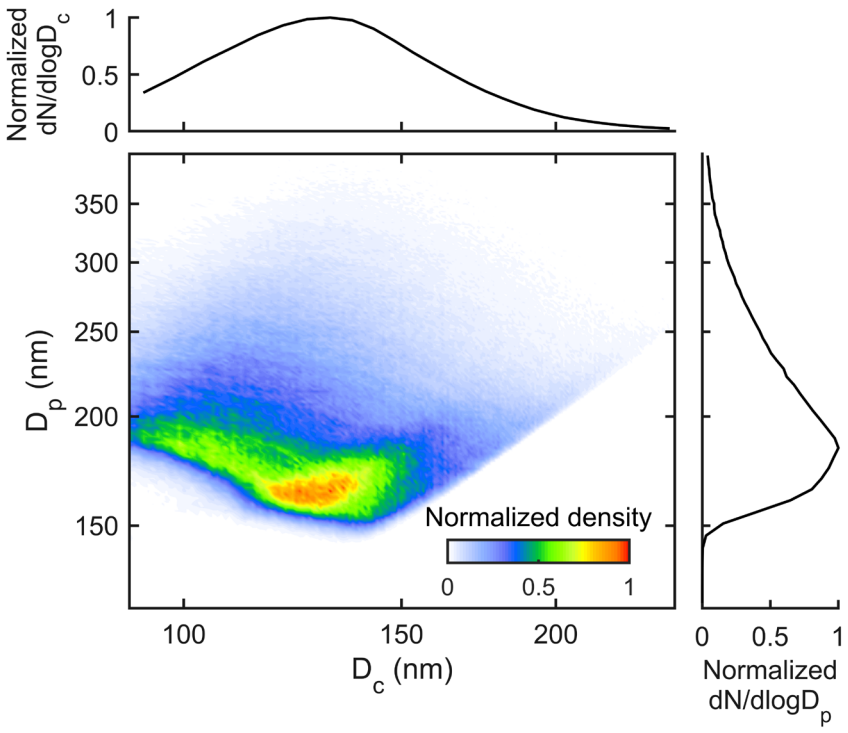


**Figure 8.** Distribution of $D_p$ and $D_c$ of internally mixed BC derived based on the BC mixing state inversion model. The image plot is a two-dimensional histogram where the color represents the number of particles falling within a specific size range, normalized to the maximum value. The number distribution of $D_c$ and $D_p$ are normalized to the peak value respectively.

Furthermore, we can derive comprehensive statistical results for BC-containing particles by utilizing the BC particle size results along with the SP2 sampling data. The statistical analysis of various physical properties of BC-containing particles in April 2022 is shown in Fig. 9. The variation of refractory BC (rBC) mass concentration is within the range of 0.24 to 0.36 μg m$^{-3}$ (Fig. 9a), and the relative number fraction of BC-containing particles to the total number of particles ranges from 0.26 to 0.36, with an average value of 0.31 (Fig. 9b). They exhibit very similar diurnal patterns. Due to less emissions and the development of the planetary boundary layer (PBL) in the daytime, minimum values occur in the afternoon, and values remain consistently high throughout the evening. The formation of the nocturnal boundary layer would favor the accumulation of pollutants, leading to elevated rBC mass concentrations during the night and early morning hours (Zhang et al., 2020). During busy traffic periods in the morning and evening, rBC mass concentration and the relative abundance of BC-containing particles increase significantly due to traffic emissions. Figure 9c shows the diurnal variation of coating thickness (calculated by $D_p - D_c$) of internally mixed BC, with an average value of 78 nm. Although the coating thickness is generally stable throughout the day, it exhibits larger variability (as indicated by the shaded area in Fig. 9c) in the afternoon. This increased variability may be attributed to the enhanced aging of BC on certain days when atmospheric conditions are conducive to photochemical reactions. There is a pronounced decrease in coating thickness of internally mixed BC at 10:00 LT. The freshly emitted BC from morning traffic undergoes aging processes and mixes with other substances in the atmosphere. Consequently, a portion of the BC

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


transitions from external to internal mixing state. This newly internally mixed BC, having experienced a short aging period, exhibits a thinner coating. This phenomenon contributes to an overall reduction in the mean coating thickness of the BC population. After 21:00 LT, the coating thickness increases gradually resulting from nighttime ageing process.

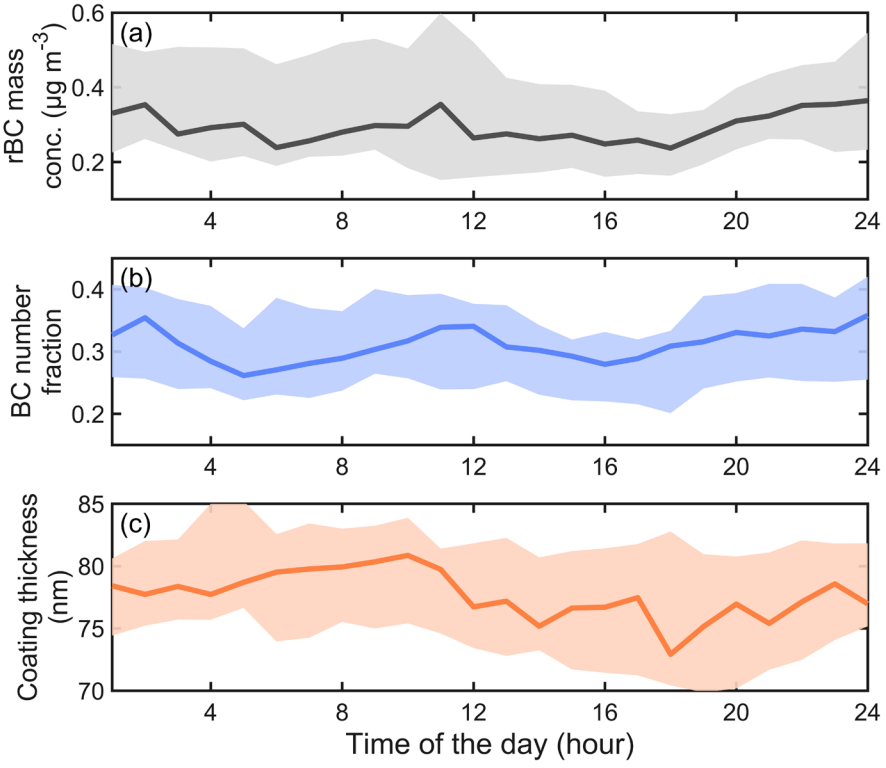


**Figure 9.** The diurnal cycles of (a) the rBC mass concentration; (b) the relative number fraction of BC-containing particles to the total number of particles; (c) the coating thickness of internally mixed BC. The solid lines represent the median value. The upper and lower boundary of the shaded area stand for the 75$^{th}$ and 25$^{th}$ percentiles, respectively.

## 5 Conclusion

This study conducted a series of explorations on the relationship between SP2 data and BC mixing state, establishing a ML-based inversion model using LightGBM to link SP2 signals with particle size and optical properties. The results show that the inversion model can efficiently derive the core and particle size of different types of particles and their optical properties. The $R^2$ between the predicted value and the actual value of the model can reach 0.98 or higher. This model can serve as a substitute for traditional physical inversion processes, simplifying the quantitative analysis process of particle size and optical properties.

Further, we employed the SHAP method to analyze the importance of the input features from SP2 signal in $D_p$ inversion model and explored their underlying physical significance. Compared to the LEO fitting method, the ML method utilizes a larger

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


coverage of signals encompassing the peak of scattering signal rather than the leading-edge data. This allows for more accurate capture of the diverse characteristics of particles. Moreover, ML method uses signals with high signal-to-noise ratio, providing better noise resistance. Additionally, by using the LightGBM algorithm, our method determines the model output by calculating the average value of the samples within the leaf nodes, further ensuring the robustness of the model.

Based on the model we have established, we can extract statistical features of BC-containing particles, including rBC mass concentration, coating thickness of internally mixed BC, BC number fraction, etc. These characteristics contribute to a better understanding of the physical properties of BC. The SP2 dataset in SORPES station was employed to validate the effectiveness and applicability of our model. The results indicate that the model can rapidly and accurately derive various physical properties of BC-containing particles. The BC number fraction varies between 0.26 and 0.36, with an average of 0.31. The diurnal variation in the coating thickness of internally mixed BC is generally stable, with an average of 78 nm. With this model, online real-time mixing state analysis of single-particle measurement is realized. The method is simple and feasible, can be widely used in environmental and climate studies.

**Code and data availability.**

The data and codes related to this article are available upon request from the corresponding author.

**Supplement.**

The supplement related to this article is available online at:

**Author contributions.**

JianW and JiapW designed and directed the study. ZT contributed to algorithm development and data analysis and wrote the manuscript. JiapW and JinW provided support for data collection. YJ, ZZ, SS, and BW helped modify the grammar of the manuscript. JianW, JiapW, CL, WN, XH, and AD contributed to the data interpretation and review of the manuscript.

**Competing interests.**

The authors declare that none of the authors has no conflict of interest.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


**Disclaimer.**

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

**Acknowledgements.**

We acknowledge the High Performance Computing Center of Nanjing University of Information Science & Technology for their support of this work.

**Financial support.**

This work was supported by the National Natural Science Foundation of China 42075098 (Jiandong Wang) and the National Key R&D Program of China (2022YFC3701000, Task 5, Jiandong Wang).

**References**

Bond, T. C. and Bergstrom, R. W.: Light Absorption by Carbonaceous Particles: An Investigative Review, Aerosol Science and Technology, 40, 27–67, https://doi.org/10.1080/02786820500421521, 2006.

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser, J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S.: Bounding the role of black carbon in the climate system: A scientific assessment, JGR Atmospheres, 118, 5380–5552, https://doi.org/10.1002/jgrd.50171, 2013.

Cappa, C. D., Onasch, T. B., Massoli, P., Worsnop, D. R., Bates, T. S., Cross, E. S., Davidovits, P., Hakala, J., Hayden, K. L., Jobson, B. T., Kolesar, K. R., Lack, D. A., Lerner, B. M., Li, S.-M., Mellon, D., Nuaaman, I., Olfert, J. S., Petäjä, T., Quinn, P. K., Song, C., Subramanian, R., Williams, E. J., and Zaveri, R. A.: Radiative Absorption Enhancements Due to the Mixing State of Atmospheric Black Carbon, Science, 337, 1078–1081, https://doi.org/10.1126/science.1223447, 2012.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L.: Machine learning and the physical sciences, Rev. Mod. Phys., 91, 045002, https://doi.org/10.1103/RevModPhys.91.045002, 2019.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., and Zeng, W.: Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agricultural Water Management, 225, 105758, https://doi.org/10.1016/j.agwat.2019.105758, 2019.

Fierce, L., Onasch, T. B., Cappa, C. D., Mazzoleni, C., China, S., Bhandari, J., Davidovits, P., Fischer, D. A., Helgestad, T., Lambe, A. T., Sedlacek, A. J., Smith, G. D., and Wolff, L.: Radiative absorption enhancements by black carbon controlled by

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


particle-to-particle heterogeneity in composition, Proc. Natl. Acad. Sci. U.S.A., 117, 5196–5203, https://doi.org/10.1073/pnas.1919723117, 2020.

Fuller, K. A., Malm, W. C., and Kreidenweis, S. M.: Effects of mixing on extinction by carbonaceous particles, J. Geophys. Res., 104, 15941–15954, https://doi.org/10.1029/1998JD100069, 1999.

Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., and Zhu, X.: Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River, JMSE, 9, 496, https://doi.org/10.3390/jmse9050496, 2021.

Gao, R. S., Schwarz, J. P., Kelly, K. K., Fahey, D. W., Watts, L. A., Thompson, T. L., Spackman, J. R., Slowik, J. G., Cross, E. S., Han, J.-H., Davidovits, P., Onasch, T. B., and Worsnop, D. R.: A Novel Method for Estimating Light-Scattering Properties of Soot Aerosols Using a Modified Single-Particle Soot Photometer, Aerosol Science and Technology, 41, 125–135, https://doi.org/10.1080/02786820601118398, 2007.

Jacobson, M. Z.: Strong radiative heating due to the mixing state of black carbon in atmospheric aerosols, Nature, 409, 695–697, 2001.

Jordan, M. I. and Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects, Science, 349, 255–260, https://doi.org/10.1126/science.aaa8415, 2015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 2017.

Liakos, K., Busato, P., Moshou, D., Pearson, S., and Bochtis, D.: Machine Learning in Agriculture: A Review, Sensors, 18, 2674, https://doi.org/10.3390/s18082674, 2018.

Liu, D., Allan, J. D., Young, D. E., Coe, H., Beddows, D., Fleming, Z. L., Flynn, M. J., Gallagher, M. W., Harrison, R. M., and Lee, J.: Size distribution, mixing state and source apportionment of black carbon aerosol in London during wintertime, Atmospheric Chemistry and Physics, 14, 10061–10084, 2014.

Liu, D., Whitehead, J., Alfarra, M. R., Reyes-Villegas, E., Spracklen, D. V., Reddington, C. L., Kong, S., Williams, P. I., Ting, Y.-C., and Haslett, S.: Black-carbon absorption enhancement in the atmosphere determined by particle mixing state, Nature Geoscience, 10, 184–188, 2017.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, http://arxiv.org/abs/1802.03888, 6 March 2019.

Matsui, H., Hamilton, D. S., and Mahowald, N. M.: Black carbon radiative effects highly sensitive to emitted particle size when resolving mixing-state diversity, Nat Commun, 9, 3446, https://doi.org/10.1038/s41467-018-05635-1, 2018.

Moteki, N. and Kondo, Y.: Effects of Mixing State on Black Carbon Measurements by Laser-Induced Incandescence, Aerosol Science and Technology, 41, 398–417, https://doi.org/10.1080/02786820701199728, 2007.

Moteki, N. and Kondo, Y.: Method to measure time-dependent scattering cross sections of particles evaporating in a laser beam, Journal of Aerosol Science, 39, 348–364, https://doi.org/10.1016/j.jaerosci.2007.12.002, 2008.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


Moteki, N., Ohata, S., Yoshida, A., and Adachi, K.: Constraining the complex refractive index of black carbon particles using the complex forward-scattering amplitude, Aerosol Science and Technology, 57, 678–699, https://doi.org/10.1080/02786826.2023.2202243, 2023.

Ramanathan, V. and Carmichael, G.: Global and regional climate changes due to black carbon, Nature geoscience, 1, 221–227, 2008.

Schnaiter, M., Linke, C., Möhler, O., Naumann, K. -H., Saathoff, H., Wagner, R., Schurath, U., and Wehner, B.: Absorption amplification of black carbon internally mixed with secondary organic aerosol, J. Geophys. Res., 110, 2005JD006046, https://doi.org/10.1029/2005JD006046, 2005.

Schwarz, J. P., Gao, R. S., Fahey, D. W., Thomson, D. S., Watts, L. A., Wilson, J. C., Reeves, J. M., Darbeheshti, M., Baumgardner, D. G., Kok, G. L., Chung, S. H., Schulz, M., Hendricks, J., Lauer, A., Kärcher, B., Slowik, J. G., Rosenlof, K. H., Thompson, T. L., Langford, A. O., Loewenstein, M., and Aikin, K. C.: Single-particle measurements of midlatitude black carbon and light-scattering aerosols from the boundary layer to the lower stratosphere, J. Geophys. Res., 111, 2006JD007076, https://doi.org/10.1029/2006JD007076, 2006.

Schwarz, J. P., Spackman, J. R., Fahey, D. W., Gao, R. S., Lohmann, U., Stier, P., Watts, L. A., Thomson, D. S., Lack, D. A., Pfister, L., Mahoney, M. J., Baumgardner, D., Wilson, J. C., and Reeves, J. M.: Coatings and their enhancement of black carbon light absorption in the tropical atmosphere, J. Geophys. Res., 113, 2007JD009042, https://doi.org/10.1029/2007JD009042, 2008.

Sedlacek, A. J., Lewis, E. R., Kleinman, L., Xu, J., and Zhang, Q.: Determination of and evidence for non-core-shell structure of particles containing black carbon using the Single-Particle Soot Photometer (SP2), Geophysical Research Letters, 39, 2012GL050905, https://doi.org/10.1029/2012GL050905, 2012.

Shiraiwa, M., Kondo, Y., Moteki, N., Takegawa, N., Sahu, L. K., Takami, A., Hatakeyama, S., Yonemura, S., and Blake, D. R.: Radiative impact of mixing state of black carbon aerosol in Asian outflow, J. Geophys. Res., 113, 2008JD010546, https://doi.org/10.1029/2008JD010546, 2008.

Slowik, J. G., Cross, E. S., Han, J.-H., Davidovits, P., Onasch, T. B., Jayne, J. T., Williams, L. R., Canagaratna, M. R., Worsnop, D. R., Chakrabarty, R. K., Moosmüller, H., Arnott, W. P., Schwarz, J. P., Gao, R.-S., Fahey, D. W., Kok, G. L., and Petzold, A.: An Inter-Comparison of Instruments Measuring Black Carbon Content of Soot Particles, Aerosol Science and Technology, 41, 295–314, https://doi.org/10.1080/02786820701197078, 2007.

Stephens, M., Turner, N., and Sandberg, J.: Particle identification by laser-induced incandescence in a solid-state laser cavity, Applied optics, 42, 3726–3736, 2003.

Sun, X., Liu, M., and Sima, Z.: A novel cryptocurrency price trend forecasting model based on LightGBM, Finance Research Letters, 32, 101084, https://doi.org/10.1016/j.frl.2018.12.032, 2020.

Tarca, A. L., Carey, V. J., Chen, X., Romero, R., and Drăghici, S.: Machine Learning and Its Applications to Biology, PLoS Comput Biol, 3, e116, https://doi.org/10.1371/journal.pcbi.0030116, 2007.

Taylor, J. W., Allan, J. D., Liu, D., Flynn, M., Weber, R., Zhang, X., Lefer, B. L., Grossberg, N., Flynn, J., and Coe, H.: Assessment of the sensitivity of core / shell parameters derived using the single-particle soot photometer to density and refractive index, Atmos. Meas. Tech., 8, 1701–1718, https://doi.org/10.5194/amt-8-1701-2015, 2015.

https://doi.org/10.5194/egusphere-2024-2496
Preprint. Discussion started: 18 October 2024


Zhang, L., Shen, F., Gao, J., Cui, S., Yue, H., Wang, J., Chen, M., and Ge, X.: Characteristics and potential sources of black carbon particles in suburban Nanjing, China, Atmospheric Pollution Research, 11, 981–991, https://doi.org/10.1016/j.apr.2020.02.011, 2020.

Zhang, Y., Zhang, Q., Cheng, Y., Su, H., Kecorius, S., Wang, Z., Wu, Z., Hu, M., Zhu, T., and Wiedensohler, A.: Measuring the morphology and density of internally mixed black carbon with SP2 and VTDMA: new insight into the absorption enhancement of black carbon in the atmosphere, Atmospheric Measurement Techniques, 9, 1833–1843, 2016.