Response to the comments of Reviewer #2 (EGUSPHERE-2024-2496)

This manuscript presents an innovative and practical approach to deriving black carbon (BC) mixing states using a machine learning (ML) framework, specifically LightGBM. The integration of SHAP analysis for interpretability and the application of the model to real-world data significantly enhance its relevance. The study is methodologically sound, with comprehensive results demonstrating the model's robustness and applicability. With previous comments addressed, further minor clarifications, particularly in defining particle categories and expanding on error analysis, will further enhance the paper's clarity and scientific rigor. Overall, this is a strong contribution to the field and can be published after minor revisions.

Response: We appreciate the reviewer's kind effort and constructive comments. All suggested improvements have been carefully implemented in the revised manuscript. We have expanded the rationale for choosing LightGBM algorithm in the introduction section, analyzed the prediction error distribution of the BC mixing state inversion model across different particle sizes, and provided a more detailed explanation of the physical significance corresponding to the important signal feature indicated by the SHAP results. Please find our point-by-point responses listed below. The reviewer's comments are in *Italic* followed by our responses and revisions (in blue).

Main Comments:

1) Line 15-20: Briefly explain why LightGBM was chosen over other models like Random Forest or Neural Networks. Highlight its advantages for handling large datasets or nonlinear relationships.

Response: Thank you for your comment. We have expanded the rationale for choosing the LightGBM algorithm in the introduction section of the revised manuscript. Regarding the abstract, we have made a brief supplement to address this point as well.

The relevant amendments of the introduction are detailed on Lines 55 to 66 :

"As an alternative, data-driven models such as machine learning (ML) can provide a good supplement to physical process-based models. ML can efficiently capture the nonlinear relationship between inputs and outputs, and has found widespread application in various fields (Carleo et al., 2019; Jordan and Mitchell, 2015; Liakos et al., 2018; Tarca et al., 2007). ***In recent years, tree-based machine learning models have gained considerable popularity due to their extremely high computational speed, satisfactory accuracy, and interpretability (Keller and Evans, 2019; Li et al., 2022; Wei et al., 2021; Yang et al., 2020). Among these, the Light Gradient Boosting Machine (LightGBM) has shown particularly outstanding performance. As a novel distributed gradient boosting framework based on decision tree algorithms, LightGBM can extract information from data more effectively than traditional tree models, excelling in handling complex non-linear relationships and high-dimensional features (Ke et al., 2017; Liu et al., 2024; Zhong et al., 2021).***

It employs innovative techniques such as gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB), which significantly improve computational efficiency while maintaining high predictive performance (Ke et al., 2017; Sun et al., 2020). Furthermore, different from some black-box models, LightGBM maintains the interpretability characteristic of tree-based models (Gan et al., 2021; Zhang et al., 2019), which can provide decision path analysis, allowing for deeper insights into the decision-making process. Considering these advantages, LightGBM can be an ideal tool for analyzing large SP2 datasets and inverting BC mixing states."

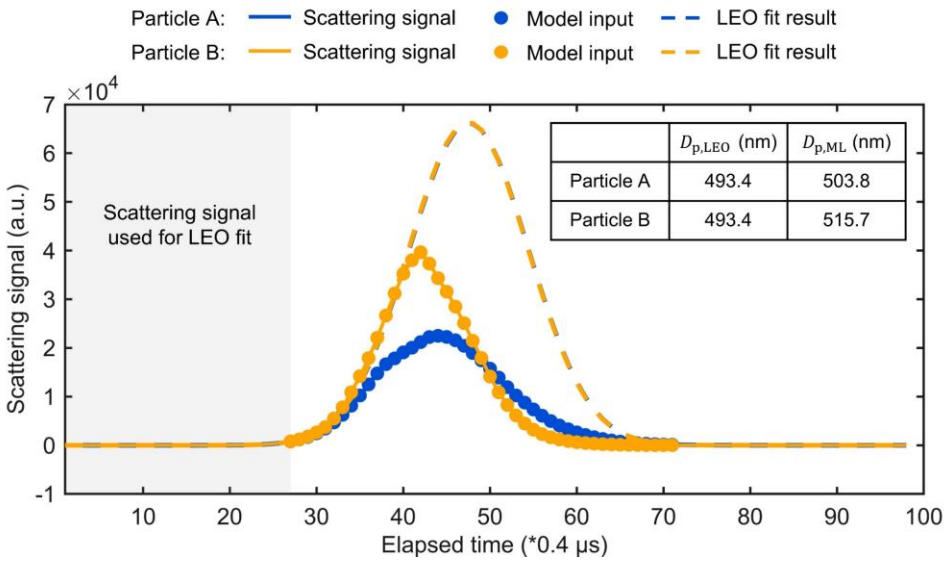
Lines 20 in abstract:

"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), ***an advanced tree-based ensemble learning algorithm****, to establish an inversion model which directly correlates SP2 signals with the mixing state of BC-containing particles."*

2) Expand the discussion on how this method improves upon or complements existing techniques, such as the LEO fitting method or other machine learning approaches. For example, what specific challenges of previous methods (e.g., noise resistance, scalability) does this model overcome?

Response: Thank you for your suggestion. A comparison between the LEO fitting results and machine learning results is shown in Fig. 9. As we discussed in the manuscript, the machine learning method utilizes more complete SP2 signals, enabling a more comprehensive characterization of particles. Furthermore, the signals used in the machine learning method have a higher signal-to-noise ratio, making it more robust against background noise compared to the LEO fitting method. The detailed discussion can be found on Lines 393 to 400 of the revised manuscript and attached below:

"Figure 9 illustrates the LEO fitting results for two different BC-containing particles. Despite nearly identical leading-edge data resulting in similar Gaussian distributions and consequently the same D_p values through LEO fitting, the complete scattering signals of these particles exhibit significant differences. The ML model, by incorporating these distinctive signal features, can effectively capture these variations, leading to different D_p predictions. Moreover, the leading edge is traditionally defined as the signal from baseline-subtracted zero up to 5 % of the maximum laser intensity (Taylor et al., 2015). As shown in Fig. 9, this portion of the signal (in the grey-shaded area) is close to the baseline, making it particularly susceptible to noise interference. Compared to LEO fitting method, the ML model utilized a broad range of signals with a high signal-to-noise ratio, demonstrating enhanced noise resistance."



*"**Figure 9.** Comparison of the scattering signal used in the D_p inversion process for internally mixed BC and corresponding calculation results from both the LEO fitting and the ML methods. The solid line represents the scattering signal obtained by SP2, and 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."*

3) Although the study uses data from a single site, could the authors discuss the expected performance of the model in different environmental conditions or geographical regions? For example, how might differences in aerosol composition affect results?

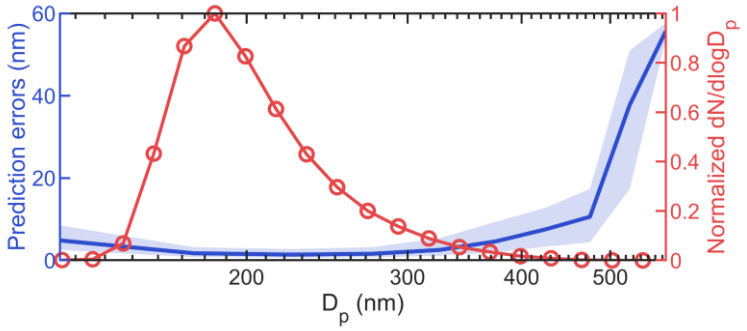
Response: Thank you for your suggestion. Operationally, SP2 data calibration is essential when conducting measurements in different regions due to variations in instrument status. Physical inversion methods inherently require these calibration procedures to ensure accuracy across different observational settings. Similarly, machine learning algorithms, including our approach, also necessitate comparable calibration processes. For instance, increased voltage leads to stronger laser intensity, resulting in enhanced SP2 scattering signals, which can cause the model to overestimate particle sizes. To address this, a voltage-related calibration function can be introduced to adjust the model's predictions. By integrating these calibrations derived from experimental data, we can enhance the model's robustness and ensure its applicability across a wide range of observational settings.

4) Consider adding uncertainty quantification for model predictions. For instance, providing confidence intervals for particle size or optical property predictions would help assess reliability in practical applications.

Response: Thank you for your comment. We have added an analysis of the prediction

errors of the BC mixing state inversion model across different particle sizes. By examining the distribution of prediction errors for the D_p of internally mixed BC in conjunction with the number size distribution (Fig. 5), we demonstrate that the model developed in this study achieves high accuracy in the 150–300 nm size range, where particles are most concentrated. Furthermore, based on the 25% and 75% percentiles of the error distribution, the model's prediction errors exhibit minimal fluctuation within this size range. We have also explained the increased prediction errors at both ends of the size distribution. This analysis helps to illustrate the model's performance across different particle size ranges and highlights its strengths and limitations. The added analysis can be found on Lines 320 to 333 in the revised manuscript and attached below:

"To comprehensively assess the model's performance across different particle size ranges, we further analyzed the prediction error distribution for D_p inversion model of internally mixed BC, as shown in Fig. 5. For particles smaller than 150 nm, the prediction errors average around 4 nm, primarily due to the low signal-to-noise ratio of their scattering signals, which introduces larger uncertainties in the LEO fitting process. The model exhibits optimal performance for particles between 150 nm and 300 nm, with an average prediction error of approximately 1.5 nm. Furthermore, based on the 25% and 75% percentiles of the error distribution, the model's prediction errors exhibit minimal fluctuation within this size range. However, prediction errors gradually increase with particle size, becoming particularly significant for particles larger than 480 nm. This trend can be attributed to occasional irregular signals at larger sizes, such as scattering or incandescence signals with abnormally broad peak widths. These signal irregularities pose challenges to the accurate characterization of particle physical properties, affecting both LEO fitting accuracy and ML model predictions, potentially leading to more pronounced discrepancies between the two methods. The number size distribution of internally mixed BC in the testing set indicates that most particles fall within the 150–300 nm range, where the model demonstrates highest accuracy. Although the prediction errors are relatively larger at both ends of the size distribution (< 150 nm and > 400 nm), the number of particles in these ranges is comparatively small, thus having limited impact on the overall performance of the model."



"Figure 5. The prediction error distribution for D_p inversion model of internally mixed BC, and normalized number size distribution for D_p of internally mixed BC in the testing set. The solid lines in error distribution represent the median value, the upper and lower boundary of the shaded area is between the 25% and 75% quantiles."

5) The SHAP analysis identifies important features in the scattering and incandescence signals. However, the physical relevance of these features (e.g., why certain regions of the signal are more predictive) could be discussed more thoroughly.

Response: Thank you for pointing this out. Through SHAP analysis, it can be observed that several crucial scattering signal features are distributed near the peak. This part of the signal represents a non-linear combination of coating evaporation and incident laser intensity changes. The pronounced signal variability within this certain region enables the machine learning model to discriminate and extract distinctive features across different particles during the training process.

Regarding the important incandescence signal features, they are primarily concentrated in the interval spanning from the initial rise of the incandescence signal to its peak intensity. The changes in the incandescence signal are closely related to the refractory BC component in BC-containing particles, thereby providing insights into the comprehensive characteristics of the entire particle.

We have further elaborated on the physical significance of the important features indicated by SHAP analysis in the revised manuscript. Specific modifications can be found on lines 367 to 371 of the revised manuscript:

"These six features, located near the peak of the scattering signal, show a positive correlation between their values and predicted D_p, as evidenced by their SHAP values. ***This part of the signal represents a non-linear combination of coating evaporation and incident laser intensity changes. Although this portion of the scattering signal deviates from the original Gaussian profile, it still correlates with the characteristics of the original particle. The intensity of the scattering signal is proportional to the particle's scattering cross-section, more pronounced signals indicate a larger scattering cross-section, and consequently, a larger D_p value.***"

Reference

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.

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.

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.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geoscientific Model Development, 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., and Geng, Y.: Application of XGBoost algorithm in the optimization of pollutant concentration, Atmospheric Research, 276, 106238, https://doi.org/10.1016/j.atmosres.2022.106238, 2022.

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, Z.-H., Weng, S.-S., Zeng, Z.-L., Ding, M.-H., Wang, Y.-Q., and Liang, Z.: Hourly land surface temperature retrieval over the Tibetan Plateau using Geo-LightGBM framework: Fusion of Himawari-8 satellite, ERA5 and site observations, Advances in Climate Change Research, 15, 623–635, https://doi.org/10.1016/j.accre.2024.06.007, 2024.

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.

Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., Li, R., and Cribb, M.: Himawari-8-derived diurnal variations in ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM), Atmospheric Chemistry and Physics, 21, 7863–7880, https://doi.org/10.5194/acp-21-7863-2021, 2021.

Yang, L., Xu, H., and Yu, S.: Estimating PM2.5 concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance, Journal of Environmental Management, 272, 111061, https://doi.org/10.1016/j.jenvman.2020.111061, 2020.

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., and Huang, L.: A Predictive Data Feature Exploration-Based Air Quality Prediction Approach, IEEE Access, 7, 30732–30743, https://doi.org/10.1109/ACCESS.2019.2897754, 2019.

Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and Zhang, W.: Robust prediction of hourly PM2.5 from meteorological data using LightGBM, National Science Review, 8, nwaa307, https://doi.org/10.1093/nsr/nwaa307, 2021.