Inversion Algorithm of Black Carbon Mixing State Based on Machine Learning

The mixing states of black carbon are widely measured by the single-particle soot photometer (SP2) instrument. This study employed a machine learning (ML) based method, light gradient boosting machine (LightGBM), to process the scattering and incandescence signals of SP2 and to retrieve the mixing states of particles. ML based method performs more efficiently than the traditional Leading-Edge-Only (LEO) approach with quite consistent retrieval outcomes. The relative importances of the selected signal features in retrieving the particle microphysical properties were studied by SHapley Additive exPlanation (SHAP) method. The authors stated that this ML based method has the potential to be a reliable noise-resistant approach to analyze the SP2 data. My major comments are attached below:

Main comments:

1. Fig.3, 4, & 5 compare the predicted particle microphysical properties with those named “actual values”. How did you define the “actual values”, and how they were obtained? Were they the outputs from LEO approach or did you use any other particle sizer instruments to measure the “actual values” of particle size? According to the meaning of “actual”, this value should be regarded as the ground-truth, or more reliable measurements of the particle size.
2. Fig. 7 shows the robustness of the ML-based retrievals by comparing the retrieved particle size with the one obtained by LEO approach. LEO method, because it utilizes part of the scattering signal, has the possibility to mischaracterize the particles with different sizes. Comments: I agree that the retrieval accuracy will be improved with more observational constraints or signal feature inputs. However, the reason that LEO method only utilizes the threshold portion but not the entire scattering signals is that the loss or vaporization of particle coatings happens after the particle started to absorb laser energy in SP2. The entire scattering signal function doesn’t reflect the scattering properties of the original mixing states of BC-containing particles (coating thickness, BC core size, etc.). Therefore, LEO method utilizes the threshold portion of the signal when the particle properties (size) doesn’t change significantly yet in SP2. Though the proposed ML-based method utilizes more signal feature, it doesn’t necessarily reflect the true size of the original coated particles.

3. Table 2: This table shows the hyperparameters for each particle type. What are the “Internally mixed BC” and “BC-containing particle” by definition? BC-containing particle is a subset of internally mixed BC in my opinion. Then why did you use different hyperparameters for them?

Line-by-line comments:

Line 73: “quality of the refractory components”, what does “quality” mean here?

Line 76: Add reference here (Gao, R. S., et al., 2007, Aerosol Science & Technology)

Line 124: “45-dimensional scattering signals”. What is the relationship between the 45-dimensional signals and the abovementioned “100-dimensional signals of SP2 data” in line 94, and 90-dimenional feature data in line 137. It would be better to provide additional contexts/descriptions of the principles of signal feature selection for optical retrievals.

Line 191: What is “GridSearchCV function”

Line 191: Please check the grammar of the sentence starting with “Based on”. What is the subject of this sentence?