Response to Reviewer #1

We sincerely thank the reviewer for the thorough, insightful, and constructive comments. The detailed feedback and careful evaluation have been extremely helpful in improving the clarity, rigor, and overall quality of the manuscript.

Below we respond to each comment point by point.

All changes are highlighted in the revised manuscript.

General Comments

Reviewer Comment:

This study aims to develop an ensemble machine learning model for the retrieval of aerosol parameters, i.e., aerosol single scattering albedo (SSA), scattering asymmetry parameter (g), effective radius (r_{eff}), and fine-mode fraction (FMF), utilizing ground-based sky radiance observations from AERONET network and radiative transfer simulations. The authors combine three powerful and widely used machine learning techniques to achieve retrievals of high accuracy with decreased computational cost and without the need for a priori assumptions and constraints compared to the traditional inversion algorithms. The retrieved products are well-presented and the ML model is sufficiently evaluated. However, the model architecture is not described in detail and some information on the configuration and the data used as input in both ML and RTM models are not very clear to me.

Overall, this is a well-written manuscript that fits into the scope of AMT. I would recommend considering the publication of this manuscript after addressing the following issues.

Response:

We thank the reviewer for the positive overall assessment. In the revised manuscript, we have:

1) added a detailed description for architecture and hyperparameters of the ensemble machine learning (EML) model (Section 2.3);

(2) clarified the radiative transfer model (RTM) configuration and the role of AERONET data to explicitly exclude any data leakage (Section 2.2; Section 2.3);

(3) revised the evaluation discussion to distinguish between model performance and independent product

comparison (Section 2.4; Section 3.2).

We further appreciate the reviewer's critical comments on key issues, which prompted us to carefully reconsider and clarify several important aspects of the methodology, leading to a substantially improved manuscript.

Specific Comments

Reviewer Comment #1:

Section 2.1: It will be useful to mention all AERONET sites used in the study and also what period the dataset covers. What are the criteria (if any) for the selection of the sites and data period?

Response:

We thank the reviewer for this comment and acknowledge that the description in the original manuscript could be clearer. Firstly, it should be noted that we did not specifically screen AERONET sites based on their geographical location. Instead, we downloaded the All Points Inversion Data products for all sites up to 2024 at once. The operational status of these sites can be both active and inactive, but as long as there is valid observation, it is sufficient.

Each set of aerosol data in All Points corresponds to a complete solar direct and Almucantar diffused observation by the photometer, but these observations require manual matching by downloading the raw measurements of the photometer from the AERONET website based on the site and time. Raw solar direct observation is the measurement of direct radiation intensity by aligning a photometer with the sun, and reliable aerosol optical depth (AOD) can be quickly and directly obtained based on the signal voltage ratio. Raw Almucantar diffuse observation refers to the measurement of sky scattering radiance at multiple angles using a photometer that maintains the same zenith angle as the Sun and changes relative azimuth angles. AOD and radiance are both considered as raw measurements of the photometer, and are also inputs to traditional, numerical optimization based full physical algorithms (Dubovik and King, 2000; Dubovik et al., 2002) as well as the machine learning algorithm used in this study, with the aim of inverting more aerosol information such as SSA, g, and r_{eff}.

Most previous research generally only used AERONET's aerosol products, so our additional step is to match these products with the original observations, mainly to verify the performance of our EML-based

algorithm on raw measurements. The time range of the data is from January 1993 to December 2024, which has been clarified in the revised manuscript: *We downloaded coincident Level 2.0 AOD and aerosol inversion products from January 1993 to December 2024, along with the corresponding raw Almucantar radiance measurements, from AERONET global sites to construct a testing set of 132,067 samples.*

Reviewer Comment #2:

Page 6, line 130: What do you mean by "randomly combined"?

Response:

Thank you for your question. I assume you are referring to the "randomly combined" aerosol size distributions, refractive indices, and surface albedo. Our training set is obtained by simulating the observation mode of a photometer using RTM (Figure 1), that is, inputting the assumed aerosol scene and outputting the observations theoretically received by the photometer. For the vector RTM, a complete scattering phase matrix is also an important aerosol input. Therefore, we choose to calculate AOD, SSA and g by the T-matrix particle scattering model using the aerosol particle size distribution and complex refractive index, and input them into the RTM. The "randomly combined" means that AERONET measurements for each parameter (including all measurements for all sites) are treated separately rather than binding them together for one measurement scene. Note that the spectral dependence of AOD, SSA, g and surface albedo is indeed preserved since it is related to aerosol microphysical and surface reflectance properties.

Reviewer Comment #3:

Section 2.2: Are the values of the EML target parameters (SSA, g, r_{eff}, and FMF) derived from the RTM using the T-matrix for SSA and g, and equations (2) and (3) for r_{eff}, and FMF? Is there any information from AERONET that you use in these computations?

Response:

Thank you very much for your comment on this and the following points. We are happy to accept your suggestion to add tables that explain the RTM and EML inputs and outputs separately. I will also provide further explanation here. Figure 1 illustrates the process of simulating photometric observations using

RTM, which mainly includes two parts: aerosol particle scattering calculation and atmospheric radiative transfer calculation. Particle scattering calculation is a preparation for better describing aerosols in atmospheric radiative transfer. In the answer to the comment #2, it was mentioned that we only sampled the complex refractive index and particle size distribution parameters of aerosols at the beginning of radiative transfer simulation from the AERONET dataset (used in equation 1), and did not sample any EML target inversion parameters (SSA, g, r_{eff}, and FMF). Therefore, these target parameters need to be calculated. The particle scattering model T-matrix can provide two optical parameters, SSA and g, while r_{eff} and FMF only need to be calculated using equations (2) and (3). So, for your first question, we answer yes.

For the second question, in my understanding, the EML target parameters in the training set do not include any AERONET information. We only independently sampled the complex refractive index, aerosol size distribution parameters, and surface albedo, which will not be duplicated with the AERONET dataset. The photometric observations calculated by RTM from this will also not be duplicated with the observations of the AERONET site photometers. The training of machine learning models requires a large amount of data. We adopt a sampling strategy based on existing observation products to ensure that the aerosol parameters in the training set are physically realistic in terms of both their ranges of variability and statistical distributions.

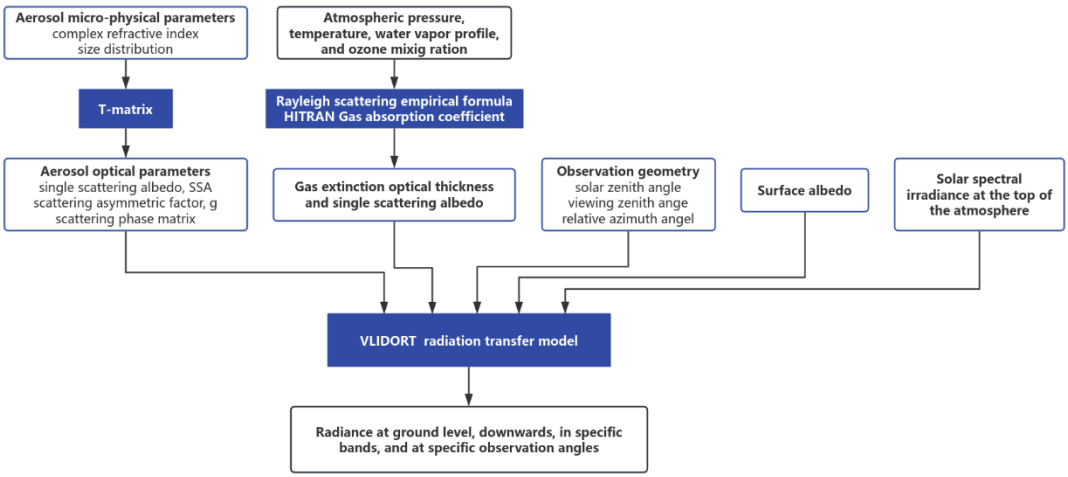

Figure 1. Forward radiation transfer calculation architecture and data. The constructed radiative transfer framework is mainly used in two aspects: firstly, simulating photometer observations under various aerosol and atmospheric scenarios to form a training set for machine learning models; secondly, verify whether the aerosol parameters inverted by the aerosol inversion algorithm can reproduce the real

observation.

Reviewer Comment #4:

Does the RTM input from AERONET include the EML target aerosol parameters (SSA, g, r_{eff}, and FMF) in any way? That could probably be considered data leakage.

Response:

Your suggestion is very important and worth discussing. The independence of the training set and data leakage are issues that machine learning must consider. Firstly, we did not directly train the model using the matched AERONET site photometer observations and official inversion products, but instead chose to use accurate RTM to simulate photometer observations and construct the training set. Secondly, instead of directly sampling the EML output target inversion parameters (SSA, g, r_{eff}, and FMF) from AERONET products, we independently sampled aerosol microphysical parameters (complex refractive index and particle size distribution) that can meet the RTM input requirements. Theoretically, there will be no identical set of data (including aerosol products and photometer observations) with existing AERONET products.

Reviewer Comment #5:

Sections 2.2 and 2.3: It is not quite clear to me if (and what) AERONET data are used as input in the RTM simulations and EML training or if AERONET data are used only for the final evaluation. The spectral AOD values used in the EML model training and cross validation are derived directly from AERONET? Apart from surface reflectance, are there any aerosol parameters from AERONET used as input in the RTM? Please clarify even if any aerosol parameters are used indirectly, e.g. to compute another parameter that is used as input. Consider adding a table with the RTM configuration and input variables (with their sources, e.g., AERONET, ERA5, climatology).

Response:

Thank you sincerely for your suggestion, and our description in the article is indeed not clear enough. Due to the rapid inversion of AOD from direct solar observations of photometry, this study aims to use diffuse sky radiation measurements to invert other properties of aerosols (SSA, g, r_{eff}, and FMF). That's why AOD is both an input for RTM and an input for EML. Existing algorithms (Dubovik and

King, 2000) also take AOD as input when inverting SSA, g, r_{eff}, and FMF. In RTM calculations and EML training, the data we use from AERONET includes aerosol parameters (spectral AOD values, complex refractive index, and particle size distribution parameters), zenith angle of photometer observation, and surface albedo. The above data were obtained through independent sampling as described above. The target parameters (SSA, g, r_{eff}, and FMF) used in EML training are indirectly calculated using T-matrix or equations (2) and (3).

The 132067 sets of Level 2.0 and 87144 sets of Level 1.5 AERONET aerosol products and corresponding photometer radiation observations mentioned in the manuscript were used to test the performance of the trained model on real observations. There are two main approaches. One is to check whether the inversion results of the EML model and AERONET product (by the official inversion algorithm of AERONET) are consistent under the same photometer radiation observation; Another is whether the aerosol parameters we reverse can physically reproduce the measurements of the photometer. Here, SSA, g, and r_{eff} from the matched AERONET data (a total of 132067+87144 sets) will be directly used, while FMF still needs to be calculated using AERONET's particle size distribution parameters and equation (3).

We accept your suggestion to add a description table for RTM configuration and input variables in the paper.

Table 1. Data and its sampling source for forward radiative transfer calculation

Variable name	Data source	Spectral dependence
Complex refractive index of aerosol	All AERONET Level 1.5 and Level 2.0 Inversion product before December 2024	yes
Aerosol size distribution parameter		no
Surface albedo		yes
Solar zenith angle/Viewing zenith angle of photometer	AERONET site photometer observation record, concentrated within the range of 50° -70°	no
Solar spectral irradiance at TOA	Climate Data Record: Solar Spectral Irradiance CDR | National Centers for Environmental Information (NCEI)	yes
Atmospheric pressure, temperature, specific humidity and ozone mixing ratio profile	ERA5 monthly mean data (2020–2024) on pressure levels	no

Reviewer Comment #6:

Section 2.3: Please refer all variables used as input in the EML model in detail. E.g., refer all AOD and

sky radiance wavelengths, all geometric parameters including all RAAs, etc. Are all 30 RAAs mentioned in line 117 used as input? Consider adding a table with all the input variables and their sources (if any), e.g., AERONET, RTM simulations etc.

Response:

We fully agree with your suggestion and have listed the input and output variables of the EML model at the beginning of Section 2. In the manuscript, we only classified and described the input variables of EML in lines 100-101, lacking a description of the order and quantity of input variables. The EML-based inversion algorithm developed in this study is a multi-band joint inversion, where wavelength dependent variables (AOD, radiance, SSA, and g) are inputted in sequence from short to long wavelengths. For Almucantar diffused sky radiation observation, the photometer measures radiation at 30 fixed relative azimuth angles (RAAs) (as stated in line 117). But when the RAA is too small, it is susceptible to direct radiation interference, so only measured radiance with a RAA greater than or equal to 7° is selected as the input.

Table 2. Input and Output variables of the EML Model

Input Variables	Count	Notes
Solar zenith angle	1	Equal to the viewing zenith angle, and the actual input is the cosine value of the angle.
Spectral AOD	4	AOD of four observation bands (440, 675, 870 and 1020 nm)
Radiance at 440nm	23	Defined at 23 relative azimuth angles (7°, 8°, 10°, 12°, 14°, 16°, 18°, 20°, 25°, 30°, 35°, 40°, 45°, 50°, 60°, 70°, 80°, 90°, 100°, 120°, 140°, 160°,180°)
Radiance at 675nm	23	Defined at 23 relative azimuth angles
Radiance at 870nm	23	Defined at 23 relative azimuth angles
Radiance at 1020nm	23	Defined at 23 relative azimuth angles
Observation geometries	23	Defined as the cosine value of the scattering angle between the incident sunlight and the observation direction of the photometer: $\cos(\theta_{sca}) = \cos(\theta_{sza})\cos(\theta_{vza}) + \sin(\theta_{sza})\sin(\theta_{vza})\cos(\theta_{raa})$. For Almucantar diffused sky radiation observations parallel to the horizontal plane, there is only one solar zenith angle and one viewing zenith angle in one scan, and the two angles are equal.

Output variables	Count	Notes
Spectral SSA	4	Single scattering albedo of aerosols in four observation bands
Spectral g	4	Scattering asymmetric factor of aerosol in four observation bands
Effective radius r_{eff}	1	Characterize the particle size of the aerosol group in the atmosphere column
Fine mode fraction FMF	1	Characterization of the volume proportion of fine particles (with a radius less than 1 micron) in the aerosol group in the atmospheric column

Reviewer Comment #7:

Page 8, line 183: Random Forest (RF), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP) are referred as “base learners”. The term “base learner” typically refers to weak and inexpensive models, such as shallow decision trees, linear models or naive Bayes models. RF and GB are ensemble powerful methods and MLPs are computationally expensive models that are often used as “strong learners” by themselves. In the context of building an ensemble of RF + GB + MLP, maybe you can say that RF, GB and MLP are used as base learners to construct a “higher-level” ensemble model, but not that these models are “base learners” by definition. Please rephrase accordingly.

Response:

We sincerely thank the reviewer for this insightful comment regarding the use of the term "base learners." We agree that Random Forest, Gradient Boosting, and Multi-Layer Perceptron are strong learners rather than weak learners. Following your suggestion, we have revised the manuscript to improve the accuracy of the terminologies. Specifically, RF, GB, and MLP are now described as first-level models within a two-level stacking ensemble framework, and their predictions are integrated by a second-level meta-learner (RidgeCV) to construct the final retrieval model. The corresponding text has been updated in Page 8 to avoid ambiguity: *In this study, Random Forest (RF), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP) were employed as first-level models to construct a higher-level ensemble retrieval framework. Random Forest represents a bagging approach that aggregates predictions from multiple decision trees trained on randomly sampled subsets of data and features (Breiman, 2001). In our RF model, 100 trees were constructed with a maximum depth of 20, and out-of-bag (OOB) estimation was enabled to assess generalization performance. Gradient Boosting is a boosting technique that builds weak learners sequentially, with each learner focusing on the residuals of its predecessors, which enables high predictive accuracy through iterative refinement (Ma, 2018). For our GB model, regression decision trees (CART) are employed as weak learners, with 100 boosting iterations and a learning rate of 0.01. The maximum tree depth is set to 8 to control model complexity. The Multi-Layer Perceptron is a feedforward neural network composed of multiple layers of interconnected neurons with nonlinear activation functions, offering strong fitting ability and architectural flexibility for capturing complex relationships (Hornik et al., 1989).*

Reviewer Comment #8:

Section 2.3: Since, these three architectures are strong learners on their own, have you tried to train the RF, GB, and MLP architectures separately and compare the results with the EML to see whether there is an actual improvement when these methods are combined? If there is no significant improvement, then maybe there is no need to build such a "higher-level" model ensemble.

Response:

We sincerely thank the reviewer for this important and fundamental comment. We fully agree that, since RF, GB, and MLP are strong learners individually, it is necessary to evaluate their standalone performance and compare it with the proposed stacking ensemble model to justify the construction of a

higher-level ensemble framework. In the manuscript, we mentioned that the evaluation model mainly relies on two sets of data, one is the validation set, which, like the training set, is obtained from RTM radiative transfer simulation (Figure 2 in the manuscript); The second group is the testing set, which comes from raw measurements of the photometer and AERONET's official inversion product (Figure 3 in the manuscript). Figure 2 summarizes the performance of three individual machine learning models and EML models in retrieving four types of aerosol parameters. It can be seen that the EML model consistently achieves the lower RMSE among all models, indicating better predictive performance. Specifically, for the testing set, the EML model yields RMSE values of 0.021, 0.028, 0.123, and 0.100 for SSA, g, r_{eff}, and FMF, respectively, which are lower than those of the Random Forest (0.030, 0.033, 0.143, 0.111), Gradient Boosting (0.037, 0.044, 0.152, 0.124), and Multi-Layer Perceptron (0.021, 0.028, 0.133, 0.098). In more detail, Figure 3-4 shows the performance of the RF model on two datasets. Compared with the EML model, its correlation coefficient (R) is smaller and the root mean square error (RMSE) is larger. Especially, the inversion accuracy of SSA on the validation set is not high. Figure 5-6 shows the performance of RB training and inversion alone. The RMSE is relatively large on both the validation and test sets, and the slope of the fitting line for scatter points using the least squares method is significantly less than 1. Figure 7-8 shows the inversion capability of MLP, which has the closest inversion performance to the EML model. However, MLP has high uncertainty when inverting 1020nm SSA on the validation set, and there is a significant difference in model performance between simulated radiation data and real observation data. Overall, in the simultaneous retrieval of multiple aerosol parameters, the individual model performs worse than the EML.

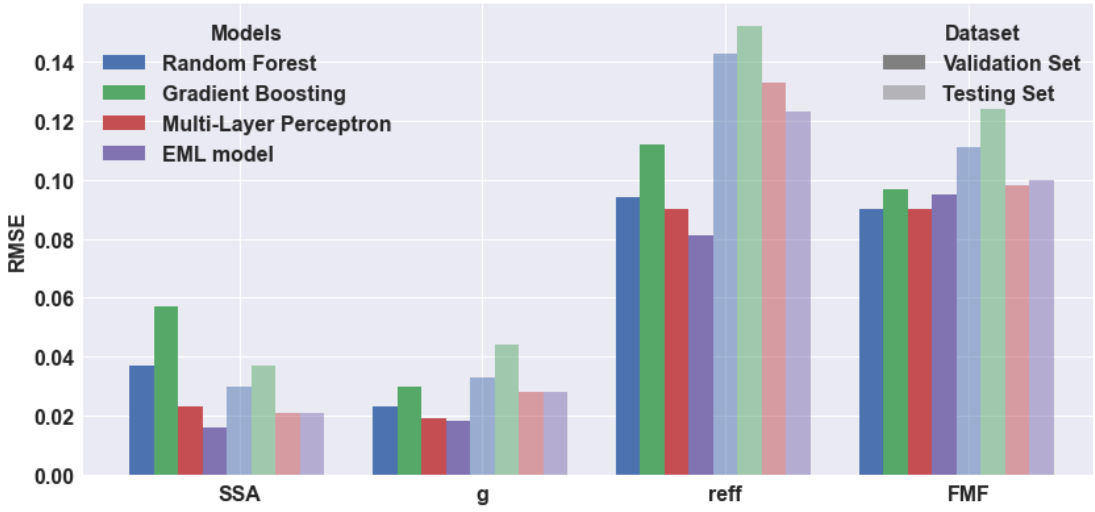

Figure 2. Root mean square error (RMSE) of three individual machine learning models and one

ensemble machine learning model (EML model) in inverting SSA, g, r_{eff}, and FMF aerosol parameters. SSA and g took the average of four observed wavelengths. The data is consistent with the scatter plots in Figures 3 to 8.

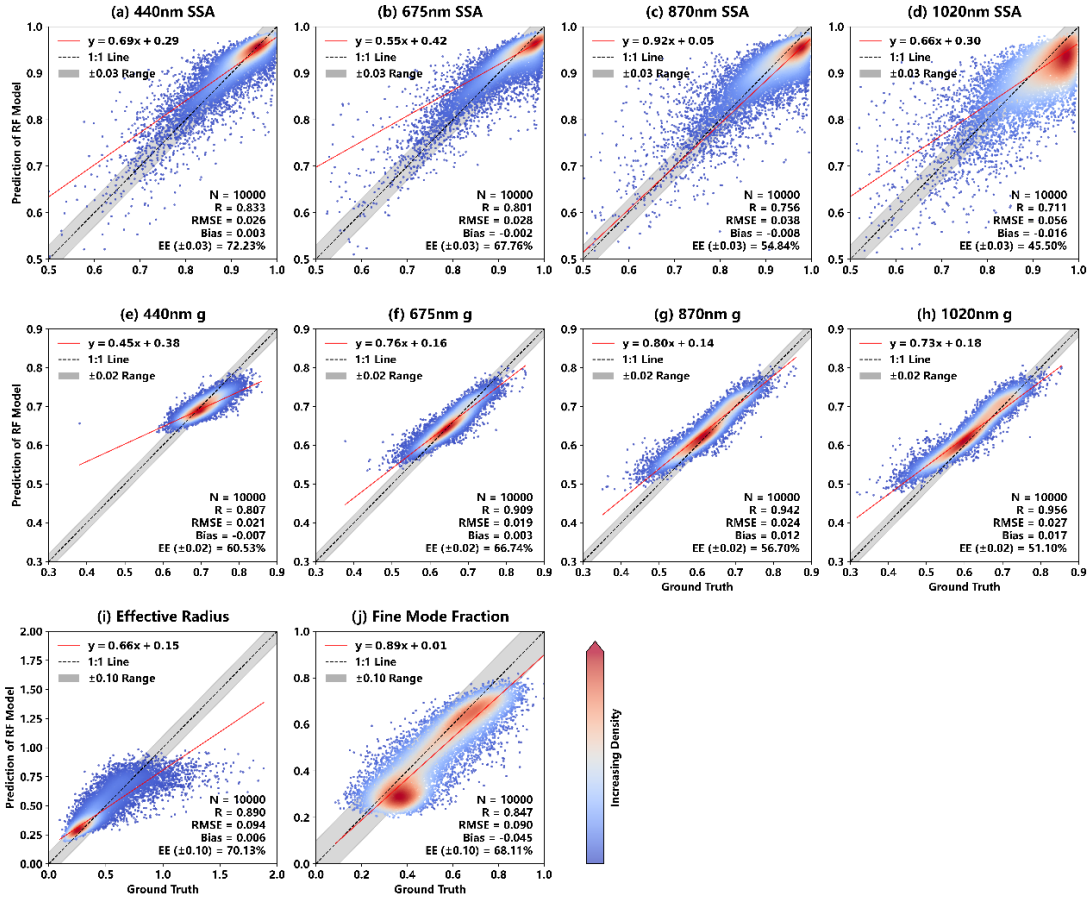

Figure 3. Aerosol parameters retrieved by the trained Random Forest model versus the ground truth on the validation set. The color of the scatter points indicates point density. Subfigures a-d correspond to retrieved variables SSA, e-h correspond to retrieved variables g, i correspond to r_{eff}, and j correspond to FMF. The four columns in the first two rows correspond to the observation bands at 440, 675, 870, and 1020 nm, respectively. The gray shaded area denotes the uncertainty range, and the red solid line is the linear regression line. The bottom-right corner of each panel shows the statistical evaluation metrics, where *N* is the total number of scatter points.

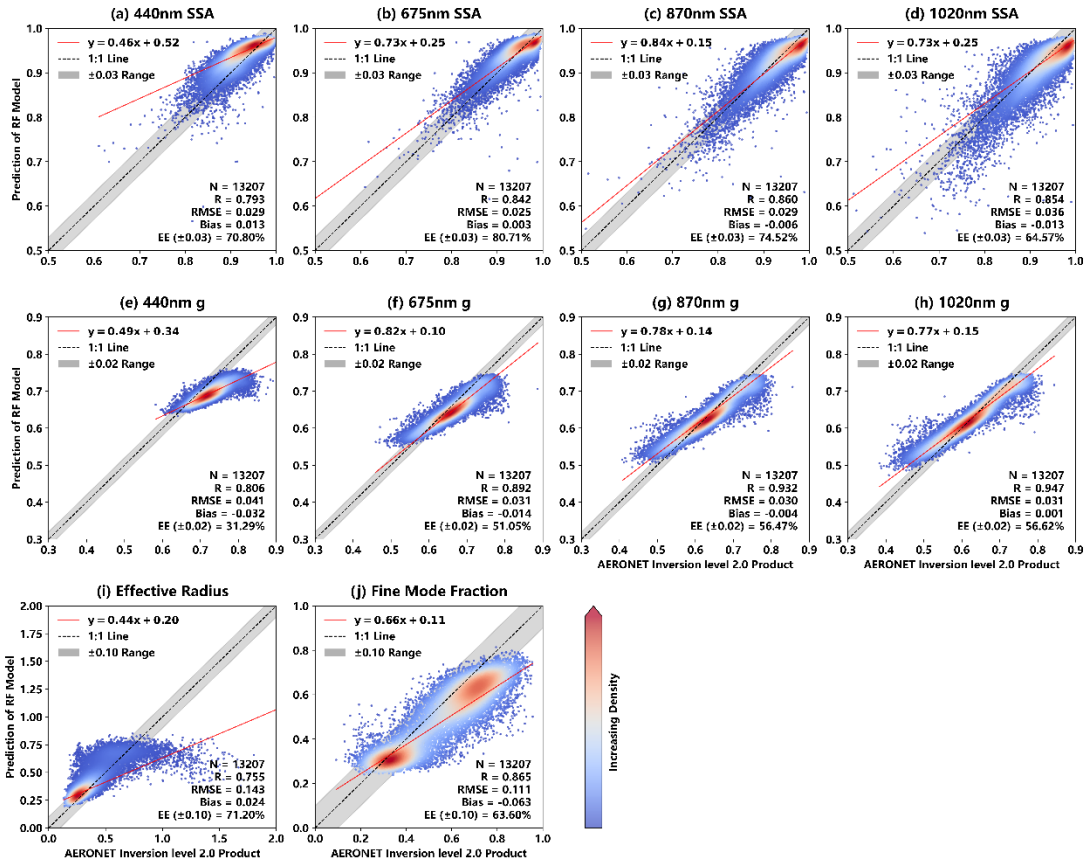

Figure 4. Aerosol parameters retrieved by the Random Forest model compared with AERONET Level 2.0 inversion products on the testing set. The plot configuration is the same as in Fig. 3. The testing set contains 132,067 raw Sun–sky photometer measurements, and the scatter points have been thinned by a factor of ten for visualization.

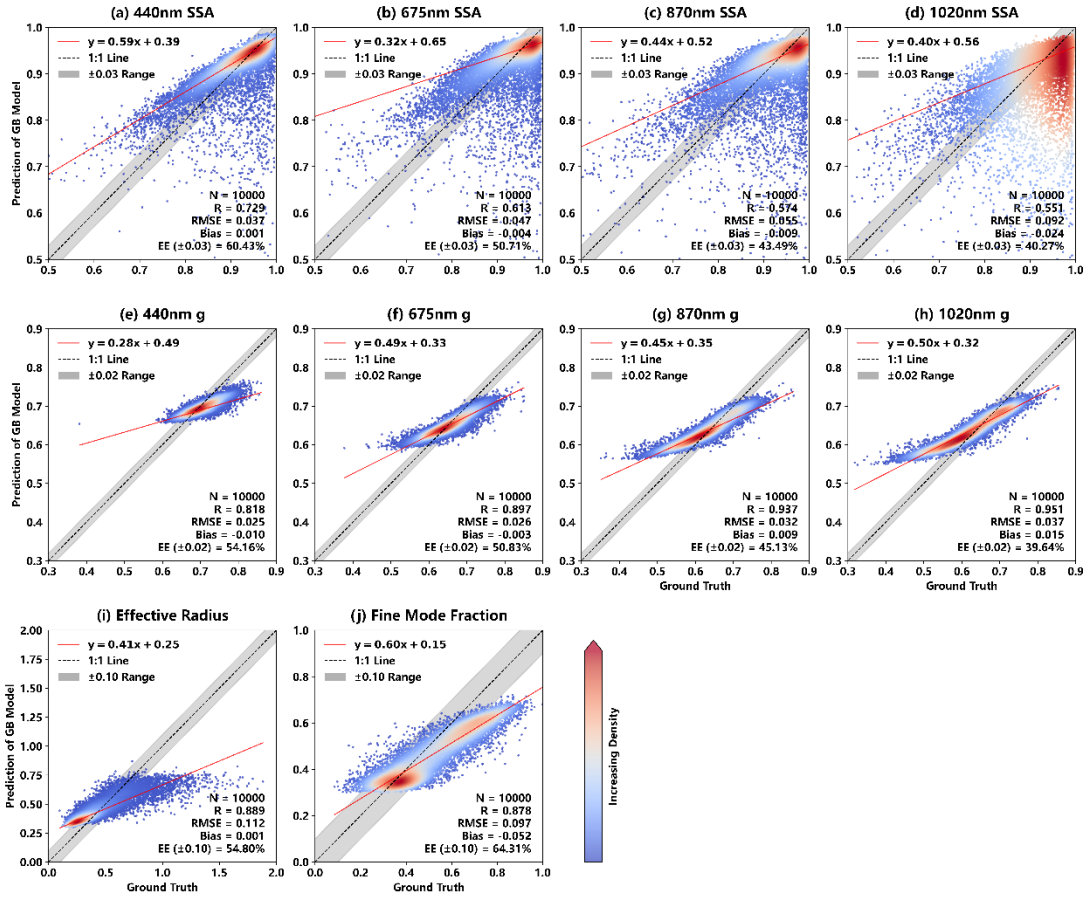

Figure 5. Aerosol parameters retrieved by the trained Gradient Boosting model versus the ground truth on the validation set. The plot configuration is the same as in Fig. 3.

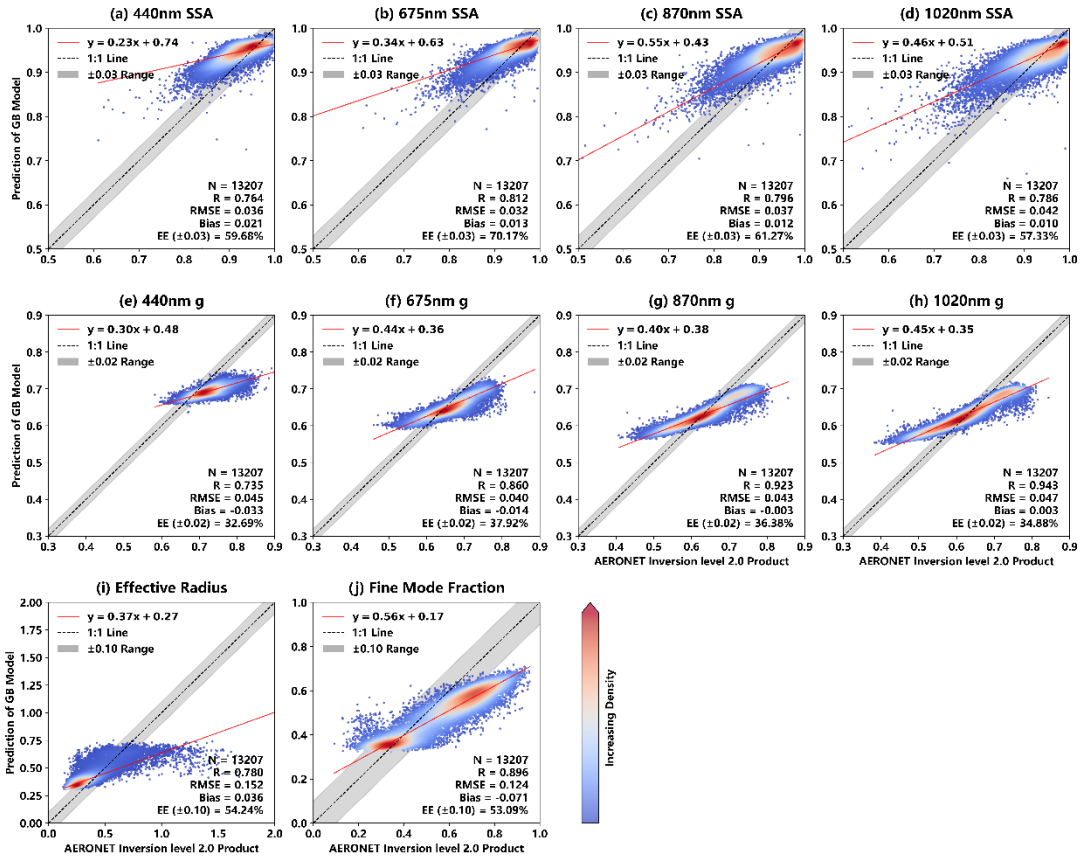

Figure 6. Aerosol parameters retrieved by the Gradient Boosting model compared with AERONET Level 2.0 inversion products on the testing set. The plot configuration is the same as in Fig. 3.

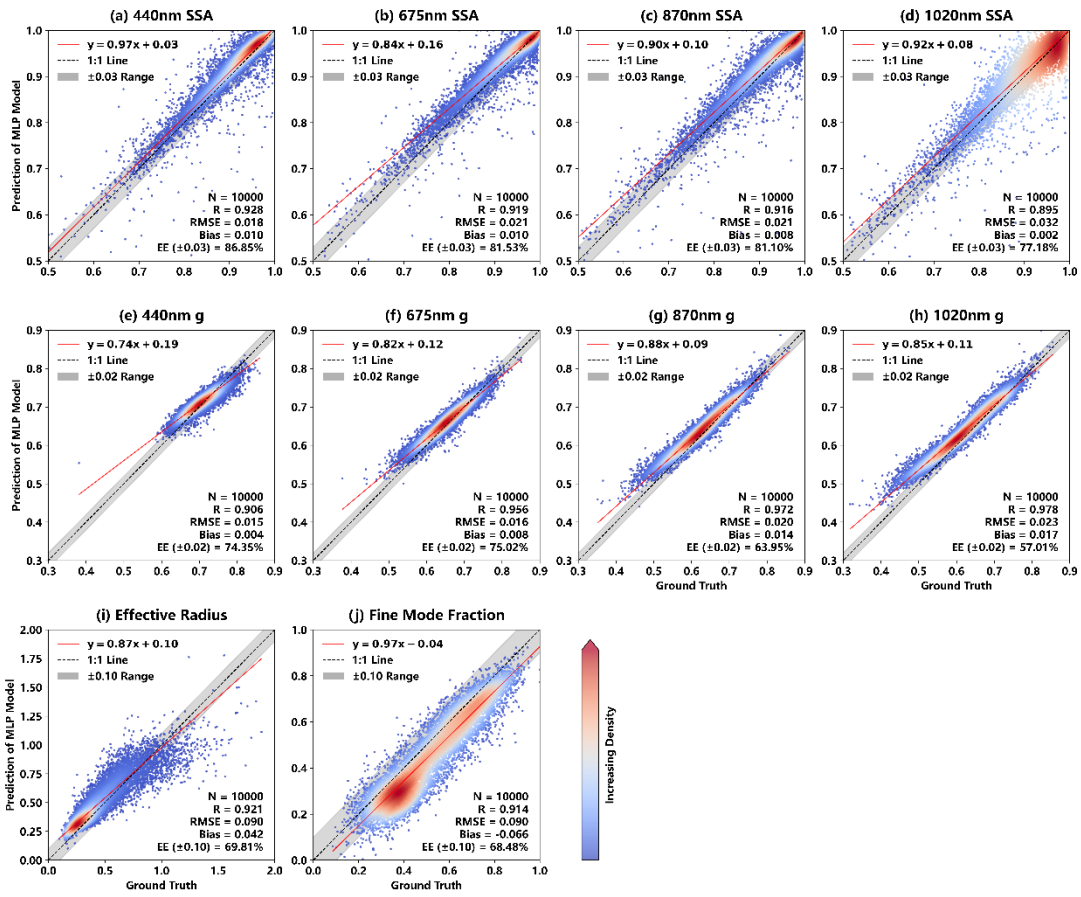

Figure 7. Aerosol parameters retrieved by the trained Multi-Layer Perceptron model versus the ground truth on the validation set. The plot configuration is the same as in Fig. 3.

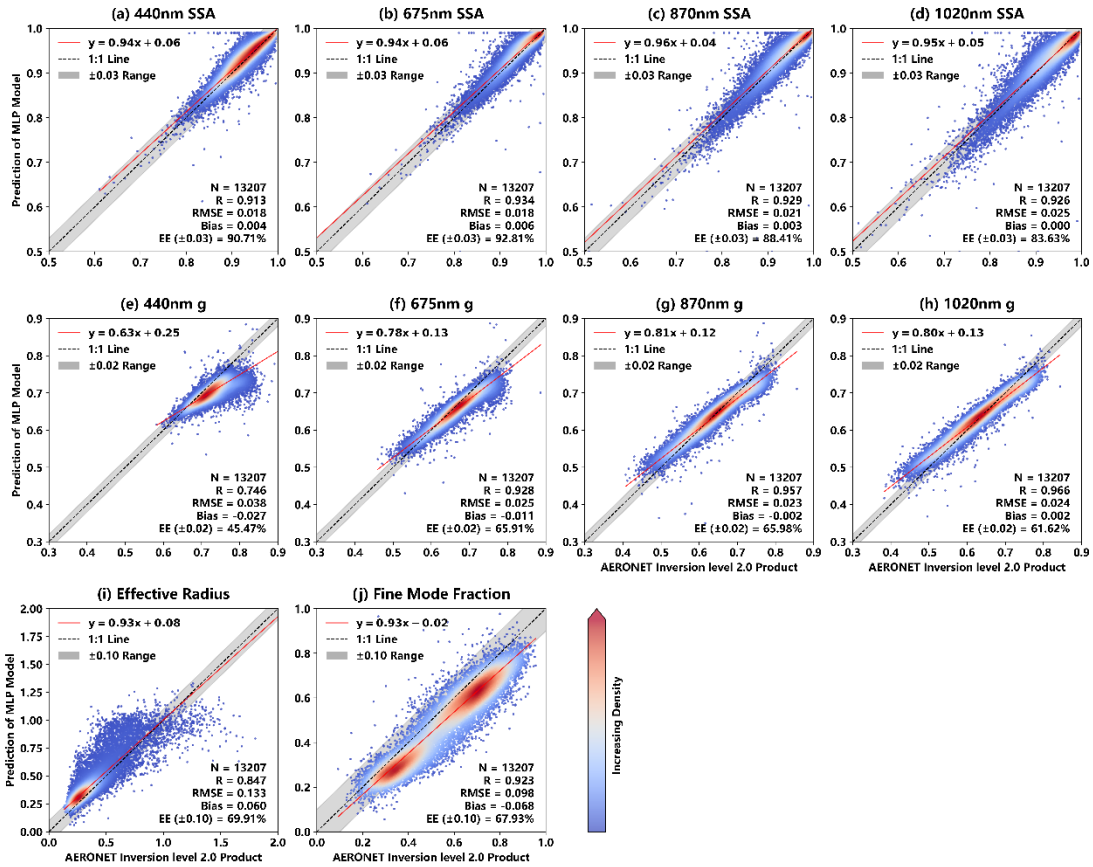

Figure 8. Aerosol parameters retrieved by the Multi-Layer Perceptron model compared with AERONET Level 2.0 inversion products on the testing set. The plot configuration is the same as in Fig. 3.

Reviewer Comment #9:

Section 2.3: What method is used as base learner in the GB model? Decision tree or other? Please clarify in the manuscript.

Response:

Thank you for your professional opinions on this and the following three, which made us realize the necessity of clarifying and understanding the model architecture. We will answer each of your questions here and clarify these details in the revised manuscript in Section 2.3. In the implemented Gradient Boosting model (scikit-learn's GradientBoostingRegressor), the base learners are regression decision trees (CART). Each boosting iteration fits a regression tree to the residuals of the previous ensemble, and the tree depth is controlled by the max_depth parameter (set to 8 in this study). The revised manuscript has added an explanation for the GB model: *For our GB model, regression decision trees (CART) are employed as weak learners, with 100 boosting iterations and a learning rate of 0.01. The maximum tree*

depth is set to 8 to control model complexity.

Reviewer Comment #10:

Section 2.3: Please mention the values used for important hyperparameters of RF and GB, e.g., number of estimators, learning rate, maximum depth, and maximum features.

Response:

In this study, the RF model comprises 100 trees with a maximum depth of 20, and out-of-bag (OOB) estimation is employed to evaluate generalization performance. For the GB model, regression decision trees (CART) are employed as weak learners, with 100 boosting iterations and a learning rate of 0.01. The maximum tree depth is set to 8 to control model complexity. The revised manuscript has added an explanation for the RF model: *In our RF model, 100 trees were constructed with a maximum depth of 20, and out-of-bag (OOB) estimation was enabled to assess general performance.*

Reviewer Comment #11:

Section 2.3: Please describe the MLP architecture and hyperparameters used.

Response:

The revised manuscript has added an explanation for MLP model: *This MLP model consists of five hidden layers (54-100-54-32-16 neurons), with a learning rate of 0.0001 and L2 regularization ($\alpha = 0.01$) to enhance training stability and prevent overfitting.*

Reviewer Comment #12:

Section 2.3: How are the different model components (MLP, RF, GB) ensembled? Do you use model stacking or other methodology? Please clarify in the manuscript.

Response:

Yes, we use a stacking strategy in the entire EML model. The specific description in the main text of the manuscript is as follows: *For the entire EML model, the predictions generated by these first-level models are used as input features for a second-level meta-learner. Specifically, a Ridge regression model with cross-validated regularization (RidgeCV) is employed to learn the optimal linear combination of the first-level predictions. This stacking strategy enables the ensemble model to adaptively weight the contributions of RF, GB, and MLP, thereby improving the model' overall retrieval performance and*

generalization ability.

Reviewer Comment #13:

Section 2.3: Does the EML model predict all target aerosol parameters (SSA, g, r_{eff}, and FMF) at once? Please clarify in the manuscript.

Response:

Yes, our EML model predict all target aerosol parameters (SSA, g, r_{eff}, and FMF) at once. As we all know, in scikit-learn, the implemented GradientBoostingRegressor natively supports only single-target regression. Therefore, for multi-output tasks, a MultiOutputRegressor wrapper is employed to extend it to multiple outputs. This wrapper independently trains one regression model for each target variable, thereby enabling the stacking framework to handle multi-output retrieval tasks.

Reviewer Comment #14:

Page 11, line 255 and Table 1: How are these metrics aggregated across all retrieved variables? Were they derived already aggregated from the EML model or were they averaged afterwards? Are there any metrics for each predicted parameter separately? To my knowledge, the ML model can report validation metrics separately for each predicted variable. If possible, include separate metrics for each target variable. Also, consider including the standard deviation for every average value you report.

Response:

Thank you very much for this insightful and highly professional comment. Your suggestion has helped us clarify the evaluation procedure and improve the completeness of original Table 1 (Table 4 in the revised manuscript). In this study, the EML model directly reports a single validation score, which by definition corresponds to the coefficient of determination (R^2), as described in equation (8) of the manuscript. For multi-output regression, this reported score represents the mean of the R^2 values across all retrieved target variables. The metric R^2 was computed in a consistent manner across all target variables. Following your valuable suggestion, in the revised manuscript we have (1) reported separate performance metrics for each predicted target variable and (2) included the standard deviation of each metric across the ten cross-validation folds. Reporting the standard deviation will better reflect the variability and robustness of model performance and provide more informative statistical evidence than

mean values alone.

We added the following discussion in the manuscript regarding the adjusted table: *The prediction score for each fold is the determination coefficient* R^2 *between the predicted value of the trained EML model and the ground truth of the output variable. The prediction scores for all retrieved variables exhibit strong consistency across the folds. For SSA, the standard deviation of the prediction scores ranges between 0.0025 and 0.0056, whereas those for* g, r_{eff}, *and FMF range from 0.0104 to 0.0120. Such consistency demonstrates that the algorithm maintains reliable predictive capability irrespective of data partitioning, further underscoring its stability and robustness.*

We sincerely appreciate your constructive recommendation, which has significantly strengthened the rigor and clarity of our results presentation.

Table 3. Prediction Scores R^2 of EML Model via Ten-fold CV

Variable / Fold	SSA				g				r_{eff}	FMF
	440nm	675nm	870nm	1020nm	470nm	675nm	870nm	1020nm		
1	0.9672	0.9515	0.9442	0.9361	0.4808	0.5925	0.6193	0.6106	0.3110	0.3508
2	0.9660	0.9480	0.9428	0.9360	0.4599	0.5737	0.6022	0.5979	0.3269	0.3509
3	0.9662	0.9380	0.9353	0.9274	0.4609	0.5774	0.6052	0.5998	0.3246	0.3572
4	0.9633	0.9517	0.9469	0.9375	0.4672	0.5832	0.6081	0.6065	0.3482	0.3801
5	0.9663	0.9442	0.9399	0.9225	0.4734	0.5916	0.6214	0.6156	0.3250	0.3638
6	0.9706	0.9495	0.9431	0.9351	0.4433	0.5641	0.5926	0.5888	0.3281	0.3557
7	0.9674	0.9513	0.9463	0.9320	0.4548	0.5594	0.5828	0.5794	0.3144	0.3520
8	0.9631	0.9456	0.9315	0.9219	0.4802	0.5835	0.6158	0.6073	0.3477	0.3814
9	0.9713	0.9500	0.9363	0.9300	0.4594	0.5714	0.5977	0.5941	0.3311	0.3593
10	0.9680	0.9490	0.9424	0.9244	0.4638	0.5676	0.5929	0.5928	0.3200	0.3436
Average	0.9669	0.9479	0.9409	0.9303	0.4644	0.5764	0.6038	0.5993	0.3280	0.3594
Standard deviation	0.0025	0.0041	0.0048	0.0056	0.0110	0.0107	0.0120	0.0104	0.0117	0.0119

Reviewer Comment #15:

Section 3.2: If the EML model is trained using target aerosol parameters (SSA, g, r_{eff}, and FMF) that are not directly retrieved from AERONET, but instead computed using RTM, T-matrix calculations, and

equations (2) and (3), then comparisons with AERONET products should be interpreted as an evaluation of the ML-derived product against an independent dataset, rather than as a direct assessment of model performance. A true evaluation of the algorithm should be based on comparisons between predicted values and target values derived using the same methodology as the training dataset, since the ML model's performance is defined by how accurately it reproduces the specific target quantities it was trained to predict. Please clarify this distinction.

Response:

We sincerely thank the reviewer for this insightful comment, which has prompted us to carefully consider the design and evaluation strategy of our algorithm. We fully agree that a true evaluation of the algorithm should be based on comparisons between predicted values and target values derived using the same methodology as the training dataset. Accordingly, in our study we employed two sets of evaluation data. The first is the validation set (Figure 2), which contains 10,000 samples generated by the RTM in the same manner as the training set (100,000 samples) but is completely independent from the training process. The primary purpose of the validation set is to assess the machine learning model's ability to reproduce the training targets accurately. The second is an independent testing set, consisting of matched raw measurements from ground-based photometers and AERONET official products (Figure 3). The goal of this testing set is to evaluate the reliability and scientific plausibility of the EML-based retrieval algorithm when applied to real-world observation. Ideally, the model should perform consistently and accurately on both sets, demonstrating that it has been effectively trained through statistical optimization and has captured physically meaningful relationships, enabling accurate aerosol parameter retrieval. Moreover, the target aerosol parameters derived from the T-matrix calculations and equations (2) and (3) are essentially consistent with the AERONET official products, as both represent aerosol characteristics that participate in radiative transfer. In fact, if the complex refractive index and size distribution from the AERONET inversion products are input into the T-matrix calculations and equations (2) and (3), the resulting target parameters (SSA, g, r_{eff} and FMF) will match those reported by AERONET. We have tried to clarify this distinction in the revised manuscript to ensure readers can correctly interpret the validation and testing results.

Reviewer Comment #16:

Section 3.3: Have you checked also the feature importance for each individual learner (RF, GB, MLP)? It would be interesting to see whether there are differences on how the different model architectures "learn" from data.

Response:

Thank you for this valuable suggestion. In this study, our primary objective was to interpret the overall behavior of the stacked ensemble model, as the final prediction is determined by the meta-learner that integrates the outputs of the first-level models. While feature importance can be computed separately for each individual learner, their importance would reflect intermediate prediction mechanisms rather than the final decision process of the ensemble. Therefore, we focused on explaining the stacked model as a whole using SHAP values, which quantify the marginal contribution of each predictor to the final prediction.

We agree that analyzing individual learners could provide complementary insights into architectural differences. Figures 8-10 show the importance analysis results of input features for three first level models: Random Forest, Gradient Boosting, and Multi-layer perceptron. The first-level model and the stacked EML model share a common feature: the retrieval of SSA relies on both direct AOD observations and sky-scattered radiance measurements. In contrast, spectral g, r_{eff}, and FMF depend more strongly on multi-angle sky diffused radiance observations. These parameters characterize the scattering directionality and asymmetry of aerosol particles and also contain information about aerosol particle size distribution. However, for individual models integrated into the EML, the observations at a given wavelength are not always the most influential features when predicting g at that wavelength. Specifically, the 675 nm observations are not the most important for retrieving 675 nm g, and likewise the 870 nm observations are not dominant for predicting g at 870 nm. Instead, the EML model better captures the wavelength-dependent contributions of the input features.

We found that the two tree-based models (RF and GB) exhibit highly similar feature importance patterns, whereas the MLP tends to concentrate more strongly on a limited subset of features (e.g., radiance at 440 nm). This difference is likely related to the intrinsic learning mechanisms of the model architectures. Tree-based models perform recursive partitioning of the feature space and aggregate decisions over many trees. As a result, their feature importance reflects cumulative contributions from multiple splits across different feature subsets. This mechanism naturally distributes importance across correlated predictors,

leading to relatively stable and similar importance patterns for RF and GB. In contrast, the MLP learns through global nonlinear transformations optimized via gradient descent. When multiple input features are strongly correlated, the network may preferentially assign larger weights to one representative feature rather than distributing weights evenly across redundant inputs (symmetry breaking under gradient-based optimization). This can result in stronger apparent dependence on a single wavelength (e.g., 440 nm radiance), even though other wavelengths contain related information.

Thank you very much for your suggestion, which allows us to further consider the differences and complementarity between different models from the perspective of feature importance, rather than just designing models based on validation set performance.

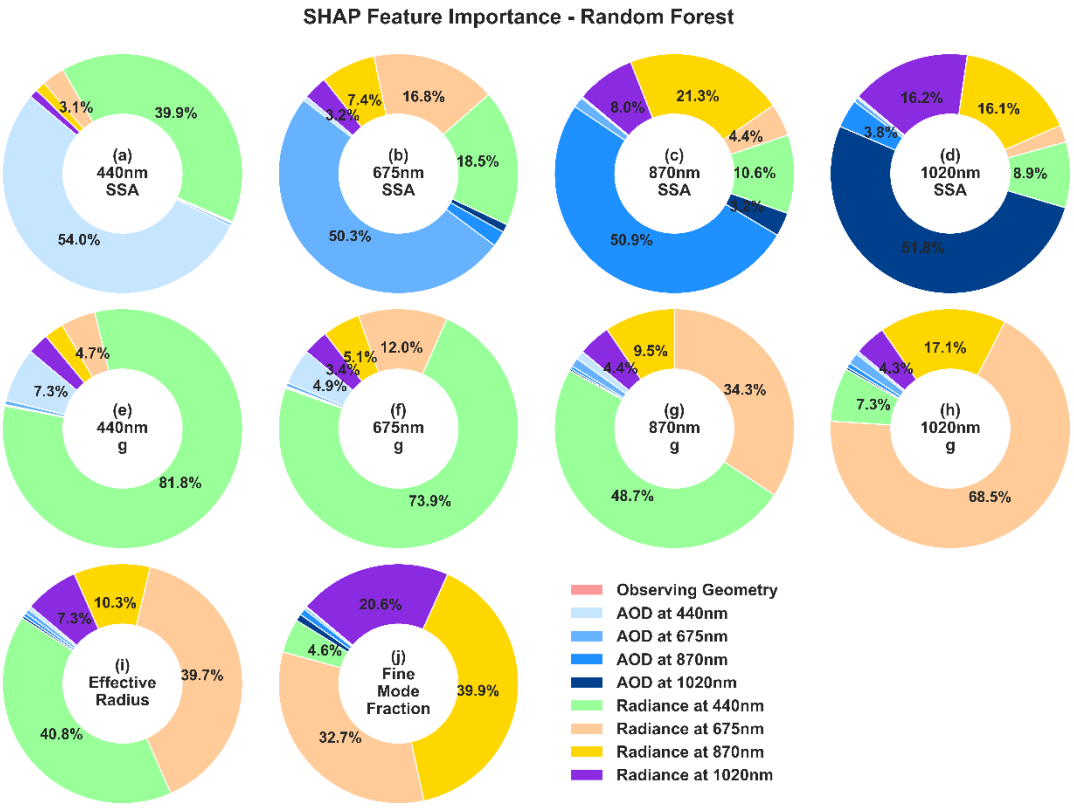

Figure 8. Importance analysis of input features for the first level model of Random Forest based on SHAP values. Subfigures a-d correspond to retrieved variables SSA, e-h correspond to retrieved variables g, i correspond to r_{eff}, and j correspond to FMF. The four columns in the first two rows correspond to the observation bands at 440, 675, 870, and 1020 nm, respectively. All 120 input features of the EML model are grouped into categories. Observation geometry includes the cosine of SZA and the scattering angle from the Almucantar scanning mode. Radiance refers to measured sky radiances

from 23 observation

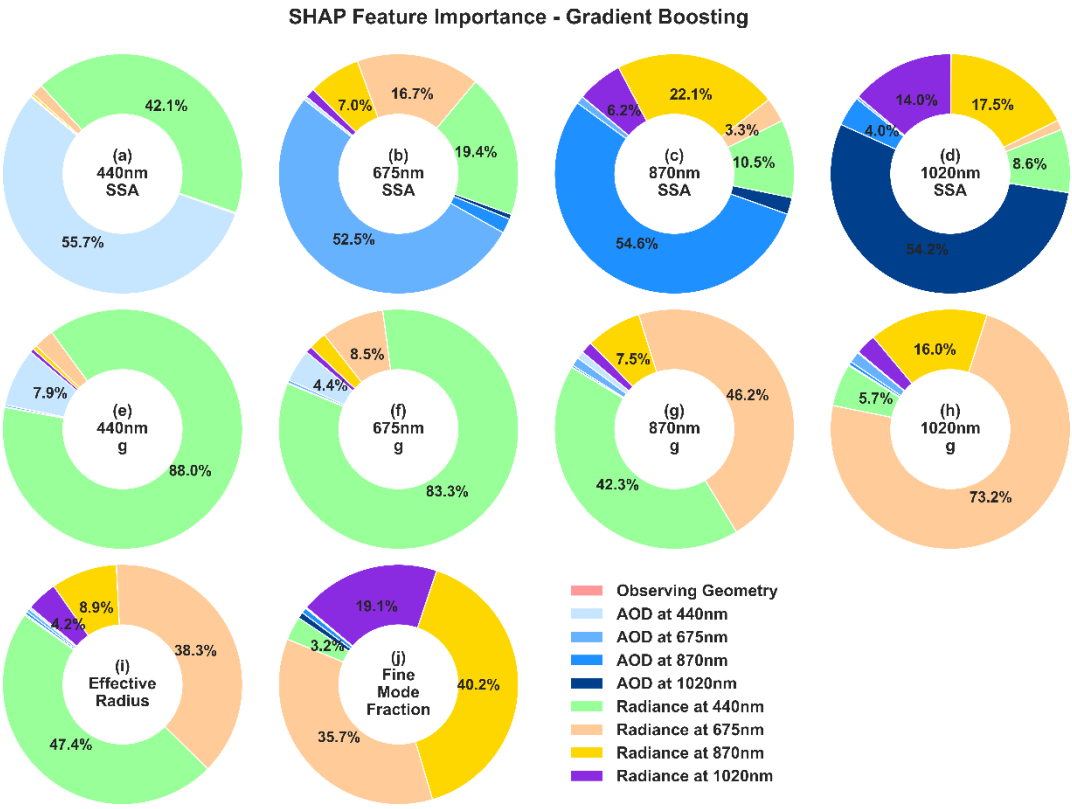

Figure 9. Importance analysis of input features for the first level model of Gradient Boosting based on SHAP values. The plot configuration is the same as in Fig. 8.

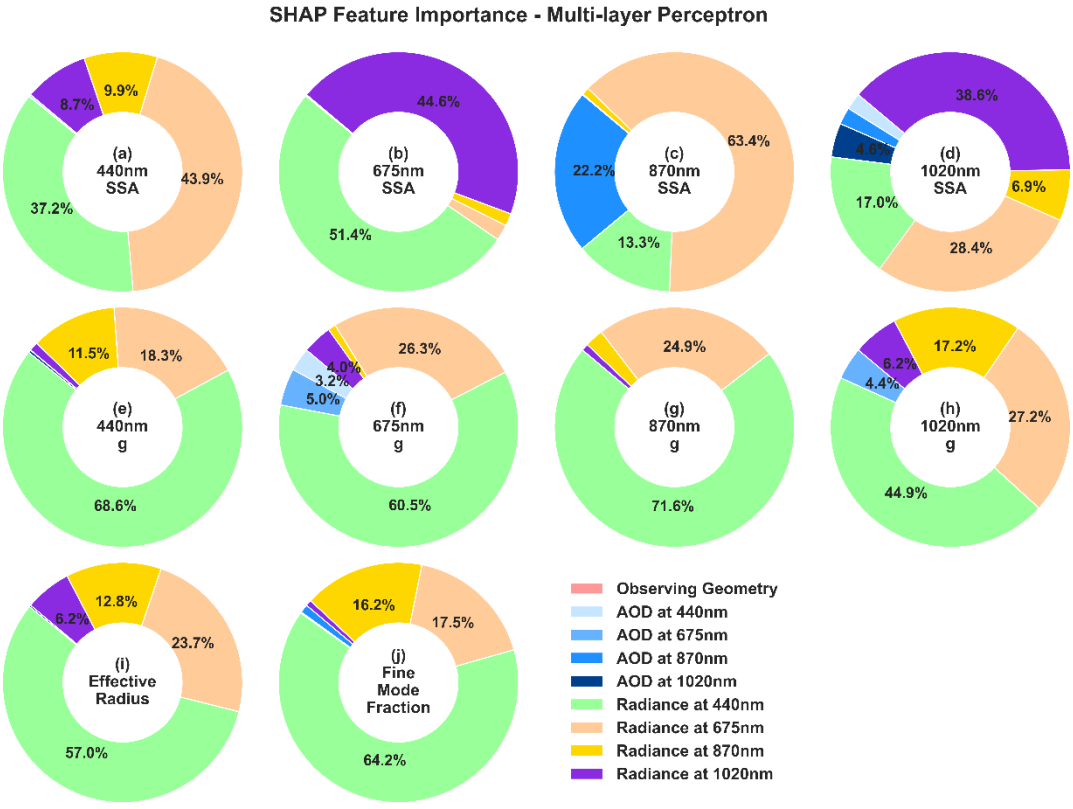

Figure 10. Importance analysis of input features for the first level model of Gradient Boosting based on SHAP values. The plot configuration is the same as in Fig. 8.

386

387 ***Reviewer Comment #17:***

388 Section 3.3: In the caption of Figure 4, it is mentioned that a number of 120 features in total are used and
389 sky radiances from 23 observation geometries are included. Please mention all features in the manuscript
390 and consider adding a table with them. Since the number of features is quite large, are there any features
391 that could be considered not important based on the feature importance analysis and could be excluded
392 from EML training? Have you also conducted an analysis to find possible correlations between the
393 different variables? By doing this, you may find further variables that could be excluded from the training.

394 ***Response:***

395 We agree that this point deserves clarification. In both this response and the manuscript, we have added
396 Table 2 to provide a detailed list of 120 variables as inputs for the EML model. As you mentioned, input
397 feature filtering is crucial for effective training of machine learning models. All input features (Table 2)
398 of the EML model have been evaluated. Removing any retained feature results in reduced performance,
399 while features with low importance have already been excluded. For example, we have indeed excluded
400 surface albedo, since ground-based observations have limited ability to constrain the surface, and the
401 proportion of ground reflection radiation in the signals received by the instrument is extremely small.
402 For a photometer, when conducting an Almucantar diffused sky radiation measurement, there are a large
403 number of observation angles. We have attempted to sparsify the input radiation, such as taking only one
404 observation every four relative azimuth angles, but the inversion performance of the model is not
405 satisfactory. This is because at different solar zenith angles (viewing zenith angles), the angle between
406 the incident light and the scattered light represented by the same relative azimuth angle is different.

407

408 ***Reviewer Comment #18:***

409 Page 16, Figure 4: Radiance at 675 nm seems to play an important role in the retrieval of g at 870 and
410 1020 nm? Do you have any clue on that?

411 ***Response:***

412 Thank you for raising this important and interesting point. This is indeed a topic worthy of discussion.

According to Mie scattering theory, the scattering properties at different wavelengths are governed by the same underlying aerosol microphysical properties, namely refractive index and particle size distribution. Radiances at 675 nm, 870 nm, and 1020 nm all constrain the same microphysical properties of aerosols. Therefore, cross-wavelength influence is physically plausible rather than unexpected. The asymmetry parameter g is a simplified representation of the particle scattering phase function, and it is strongly controlled by aerosol microphysical properties, particularly the particle size parameter $\alpha = \frac{2\pi r}{\lambda}$. In general, shorter wavelengths tend to be more sensitive to variations in particle size distribution. As particle size increases, forward scattering becomes more pronounced, leading to larger values of g. Consequently, radiance measurements at one wavelength can provide indirect constraints on the particle size distribution, which in turn affects the retrieval of g at other wavelengths. In Figure 4, radiance at 440 nm also influences the retrieval of g at neighboring wavelengths such as 675 nm and 870 nm. Similarly, the relatively strong contribution of 675 nm radiance to the retrieval of g at 870 nm and 1020 nm can be interpreted as reflecting the shared dependence on the same particle size structure. The above explanation is based on our current understanding of particle scattering and radiative transfer processes. We acknowledge that this interpretation may not be exhaustive, and further investigation is needed to better quantify and isolate the underlying mechanisms.

Reviewer Comment #19:

Page 16, Figure 4: Consider showing the importance of AOD at each wavelength separately.

Response:

We agree that it is more reasonable to display the feature importance of AOD in different wavelength, and Figure 4 has been revised in the manuscript. The importance of AOD in SSA inversion is above 40%, and distinguishing wavelengths is more conducive to analyzing the model.

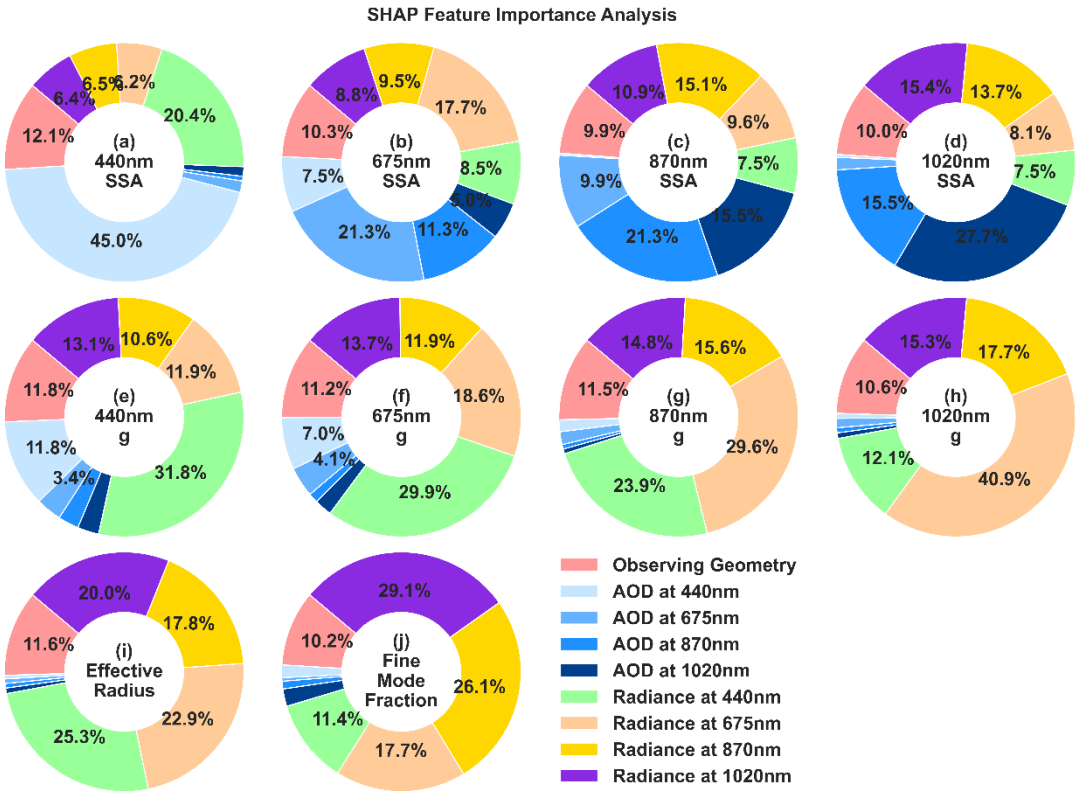

Figure 11. Importance analysis of input features based on SHAP values. Subfigures a-d correspond to retrieved variables SSA, e-h correspond to retrieved variables g, i correspond to r_{eff}, and j correspond to FMF. The four columns in the first two rows correspond to the observation bands at 440, 675, 870, and 1020 nm, respectively. All 120 input features of the EML model are grouped into categories. Observation geometry includes the cosine of SZA and the scattering angle from the Almucantar scanning mode. Radiance refers to measured sky radiances from 23 observation geometries. Values less than 3% are hidden.

Reviewer Comment #20:

Appendix B: Was the EML trained also on data with AOD < 0.4? If not, then maybe it's expected to fail such cases. To test EML performance on data corresponding to AOD < 0.4 and present meaningful results, maybe you should include such cases in the model training.

Response:

Thank you for providing such an important suggestion. The distribution of values for all input and output variables on the training set has a significant impact on model training and performance on the validation/testing set. The source of RTM input data is shown in Table 1, where aerosol samples include

both Level 1.5 and Level 2.0 AERONET Inverse products. Therefore, the EML was trained on data with AOD<0.4. The following Figure 8 shows the distribution of values for 440nm AOD on the training set. In the manuscript, the correlation coefficient R and root mean square error RMSE of the scatter points in Figure B1 and Figure B2 are not as good as those in Figure 2 and Figure 3, but the scatter points are still distributed around the 1:1 line as a whole, only relatively scattered.

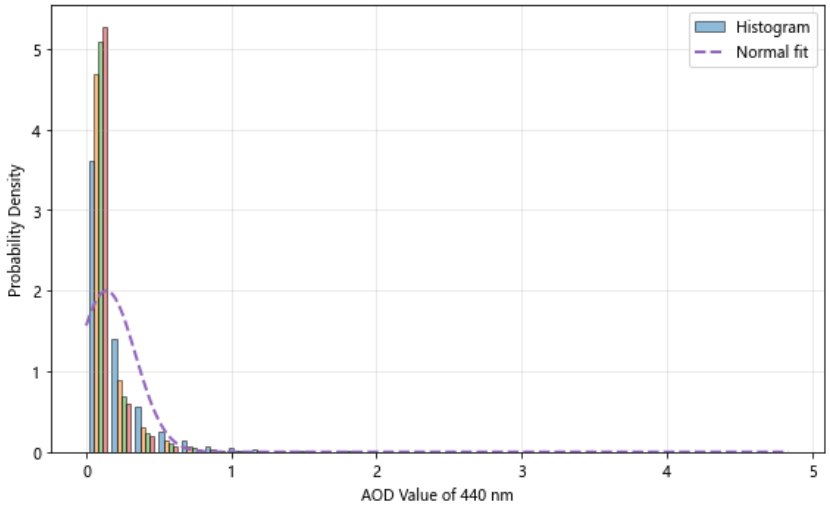

Figure 12. The distribution of values for 440nm AOD on the training set.

Technical Corrections

Reviewer Comment #21:

Page 2, line 45: The period mark after "radiative transfer model (RTM)" looks like a typing error.

Response:

We thank the reviewer for pointing out this typographical error. The extra period after "radiative transfer model (RTM)" has been removed in the revised manuscript: *The core of this algorithm is a numerical optimization process that iteratively adjusts the aerosol size distribution and complex refractive index until the observed radiance is reproduced via a radiative transfer model (RTM) (Dubovik and King, 2000; Dubovik et al., 2002).*

Reviewer Comment #22:

Page 5, Figure 1: There are some typing errors like in the words "meansurement" and "redisual" in the "Algorithm Evaluation" box of the flowchart.

Response:

We sincerely thank the reviewer for the careful reading of our manuscript and for pointing out these typographical errors. The misspellings in Figure 1 have been corrected: "meansurement" has been revised to "*measurement*" and "redisual" has been revised to "*residual*" in the "Algorithm Evaluation" box of the flowchart.

Reviewer Comment #23:

Page 2, Figure 2: There are three rows, not four. Please rephrase the following sentence in the caption accordingly: "The four rows correspond to the four retrieved variables: SSA, g, r_{eff}, and FMF."

Response:

We thank the reviewer for pointing this out. Indeed, there are only three rows of subplots. In the revised manuscript, we have updated the figure caption and now label the subplots with letters (a–j) to indicate the corresponding retrieved variables: *Subfigures a-d correspond to retrieved variables SSA, e-h correspond to retrieved variables* g*, i correspond to* r_{eff}*, and j correspond to FMF.*

Reviewer Comment #24:

Page 2, Figure 2: The sentence "The four columns represent the observation bands at 440, 675, 870, and 1020 nm." applies to the first and second row, but not the last one. Please clarify in the caption.

Response:

We thank the reviewer for the comment. Only SSA and g are spectrally dependent, and we have identified that *"The four columns in the first two rows correspond to the observation bands at 440, 675, 870, and 1020 nm, respectively".*

Reviewer Comment #25:

Page 16, Figure 4: Similarly, replace "four rows" with "three rows" in the caption and clarify that "The four columns represent the observation wavelengths…" applies only to the first and second rows.

Response:

Thank you for your comment. The caption of Figure 4 has been revised in a manner consistent with Figure 2.

Reviewer Comment #26:

Page 17, Figure 5: The colorbar needs improvement. Maybe consider removing the values from colorbar which, to my understanding, do not correspond to the metrics' values, and change the colorbar title or maybe remove the colorbar at all and keep only the explanation in the figure caption.

Response:

We fully agree with the reviewer's comment. The colorbar values do not correspond to the actual metric values and may easily cause misunderstanding, and the values themselves have no physical meaning. Following your suggestion, we have removed the colorbar and retained only the explanation in the figure caption: *The color shading beneath each number does not denote absolute metric values. Rather, lighter shades indicate better model performance for the output variable in a given row with respect to the metric in the corresponding column, while darker shades (approaching deep blue) indicate worse performance.*

Reviewer Comment #27:

Page 20, Figure 7: Please replace " the numbers inside each box " with " the numbers above each box " in the caption.

Response:

We thank the reviewer for the suggestion. The figure caption has been revised accordingly, replacing "the numbers inside each box" with "*the numbers above each box*" in the revised manuscript.

Finally, we sincerely thank the reviewer for the careful and thorough evaluation of our manuscript. The insightful comments and constructive suggestions have greatly helped us improve the quality and clarity of the paper, and have also highlighted several aspects of the algorithm design that deserve further consideration and refinement. We truly appreciate the time and effort devoted to reviewing our work.

Response to Reviewer #2

We sincerely thank the reviewer for the thorough evaluation of our manuscript and for the positive assessment of our work. The constructive comments and valuable suggestions have helped us identify several areas where the manuscript can be further improved in terms of clarity, interpretation, and technical rigor.

We have carefully addressed each comment point by point in the following responses. All corresponding revisions have been incorporated into the manuscript and are clearly marked in the revised version.

General Comments

This study presents an ensemble machine learning (EML)–based aerosol retrieval algorithm for estimating both aerosol optical and microphysical properties from ground-based Sun–sky photometer measurements. The results demonstrate performance comparable to AERONET operational products, while achieving a computational efficiency improvement of more than five orders of magnitude. The authors also clearly describe the limitations of the current approach. Overall, the manuscript is well organized, with appropriate and informative figures and tables. Therefore, this work is relevant to the AMT readership and fits well within the journal's scope. I recommend publication after minor revision.

However, the manuscript would benefit from a more in-depth interpretation of the results, rather than primarily repeating quantitative evaluation metrics that are already evident from the figures and tables. In addition, although a 5% radiance uncertainty is assumed based on earlier studies, it would be valuable to discuss whether this assumption remains valid given current state-of-the-art radiative transfer calculations and instrumental capabilities.

Response:

Thank you for your constructive and inspiring comments. When discussing algorithm inversion results, we should not only look at the data, but also analyze the physical mechanisms reflected behind it. Specifically, we have added a discussion on the possible reasons behind the ability of multi input/output models to distinguish specific wavelength observations and simultaneously utilize multi wavelength observations. The 5% radiance uncertainty adopted in this study follows commonly used assumptions in previous Sun–sky photometer retrieval studies and AERONET-related algorithms, where typical sky radiance uncertainties are reported to be on the order of 3–5%, while direct solar measurements are often

within 1–2%. With improvements in instrument calibration and radiative transfer modeling, the actual uncertainty for high-quality observations may be lower under ideal conditions. Therefore, the 5% assumption can be considered a conservative estimate. In our framework, this uncertainty level is used to ensure that the model remains robust under realistic observational noise and does not overfit idealized noise-free simulations. The 5% residual may also come from other factors involved in radiative transfer calculations, such as solar incident radiation at the top of the atmosphere, surface albedo, gas absorption, and Rayleigh scattering. The setting of observation error covariance matrix is crucial for iterative statistical algorithms based on Bayesian optimization. However, for machine learning algorithms, a 5% observation uncertainty does not affect the aerosol parameter inversion results, but only evaluates the model in radiation space. The smaller the residual between the radiation calculated by simulating the aerosol parameters using inversion and the observation, the more accurate the inversion results are. We must admit that compared to statistical algorithms, machine learning models are difficult to provide posterior errors. We estimate the propagation error through multiple perturbations (Section 2.4), and whether to choose a perturbation intensity of 5% also affects the error evaluation of the inversion results. Excessive perturbation intensity may overestimate the error, but currently there is no particularly scientific method to find a more suitable value. We have added discussion explanations in Section 3.4: *According to Section 2.4, the evaluation of propagation error depends on the intensity of perturbation to the input radiation. The stronger the perturbation, the greater the error. In the future, the accuracy of the instruments will likely improve, and we hope to achieve better accuracy in the inversion results.*

Specific Comments

Reviewer Comment #1:

220 # --> n or n* etc, otherwise please define #

Response:

Thank you for pointing out that we did miss the definition of # in the formula (11). The following discussion has been added to the main text: *# is a counting symbol representing the number of points in the subsequent set.* The condition that a set element needs to satisfy is that the absolute deviation between y and $\hat{y}$ is less than the specified uncertainty.

Reviewer Comment #2:

262 how do you make sure these results are without overfitting?

Response:

We fully agree with your point of view that preventing overfitting in model training is an issue that must be considered in machine learning model training. On one hand, our model was trained on a dataset of 100,000 sets of simulated radiative transfer patterns. During the training process, 5% of the data will not participate in the optimization of the model's structural parameters, but will be used to monitor whether the model is overfitting. After each round of training, the model needs to make predictions on this 5% of the data, and the loss function between the predicted values and the ground truth should not have a significant upward trend. If the loss function changes too little or shows a significant upward trend compared to the previous two rounds of training, the training will be terminated. On the other hand, when designing hyperparameters for the EML model, we will minimize the model complexity as much as possible. For example, the number of decision trees in the Random Forest model should not be too many, and the number of network layers and neurons in the Multi-Layer Perceptron should not be too many. Thirdly, the prediction performance of the model on training set, validation set and test set has not significantly deteriorated or improved. The validation set is comprised of simulated data that did not participate in model training, while the test set is derived from real observations from a photometer.

Reviewer Comment #3:

284-286 please suggest that these results are faster by O(10^5), for example, as stated in the Summary section.

Response:

We should indeed further clarify the magnitude of the speed increase in lines 333-335 in the revised manuscript, which has been revised in the manuscript: *It requires only 0.18 milliseconds to invert a single measurement, which corresponds to a speed improvement on the order of* 10^5*, since traditional numerical retrieval algorithms often take several minutes per case.* But we acknowledge that the speed improvement of inversion algorithms cannot be estimated by specific values, because for traditional iterative optimization algorithms, the number of iteration steps and radiation transfer calculations

required for different cases are different, usually taking minutes. For the EML-based inversion algorithm, the speed is also affected by computer computing power, but the inversion of a case is in milliseconds. Therefore, the fold increase in inversion speed is approximately on the order of 10^5.

Reviewer Comment #4:

323 For example, why 440 nm radiance has such high impact in determining SSA? With stronger Rayleigh scattering, the presence of absorbing aerosol shows larger difference…e.g. similarly for g as well, and throughout the following explanations.

Response:

Thank you very much for your suggestion. We should not only explain the data, but also increase our consideration of physical laws. Firstly, we have made adjustments to Figure 4 in the revised manuscript, mainly by refining the spectral AOD feature and displaying the importance of the AOD features for each of the four observation bands. For the inversion of 440nm SSA, the most important input variables are the AOD and radiance observed by the photometer at 440nm. AOD is obtained by direct observation with a photometer, which characterizes the total column amount of aerosols along the entire optical path, including the absorption extinction and scattering extinction of aerosols. The radiance observed in Almucantar mode is mainly sky scattered light, and the path length of scattered light will be significantly different from direct light. The absorption and scattering characteristics of aerosols in the atmosphere will significantly affect the observed radiation signal. Therefore, the radiance observed from different angles is crucial for the inversion of SSA and g.

The revised manuscript has added Tables 2 and 3, which provide a detailed list of variables that are simultaneously input or output to the EML. In Figure 4, it can be seen that the EML model effectively extracts and utilizes observation data from specific wavelength bands for aerosol parameter inversion at corresponding wavelengths. That is to say, the multi variable input-output EML model can learn wavelength sensitivity. As you said, Rayleigh scattering is stronger at shorter wavelengths, and absorbing aerosols such as black carbon and brown carbon are more sensitive in the blue light band. The sensitivity of SSA and g at 440nm to radiation at 440nm is stronger than that at longer wavelengths. The following discussion has been added in lines 376-379: *Rayleigh scattering is stronger at shorter wavelengths, and absorbing aerosols such as black carbon and brown carbon more heavily impact the blue light band. The*

114 *sensitivity of SSA and g at 440nm to radiation at 440nm is stronger than in longer wavelength bands.*

115 Another question is why the inversion parameters of one wavelength band are also affected by other

116 wavelengths. *The three parameters* g, r_{eff} *and FMF essentially reflect the particle size distribution of*

117 *aerosols, and this microphysical property will inevitably affect radiation transmission at multiple*

118 *wavelengths.*

119

120 **Table 2. Input variables of the EML Model**

Input Variables	Count	Notes
Solar zenith angle	1	Equal to the viewing zenith angle, and the actual input is the cosine value of the angle.
Spectral AOD	4	AOD of four observation bands (440, 675, 870 and 1020 nm)
Radiance at 440nm	23	Defined at 23 relative azimuth angles (7°, 8°, 10°, 12°, 14°, 16°, 18°, 20°, 25°, 30°, 35°, 40°, 45°, 50°, 60°, 70°, 80°, 90°, 100°, 120°, 140°, 160°,180°)
Radiance at 675nm	23	Defined at 23 relative azimuth angles
Radiance at 870nm	23	Defined at 23 relative azimuth angles
Radiance at 1020nm	23	Defined at 23 relative azimuth angles
Observation geometries	23	Defined as the cosine value of the scattering angle between the incident sunlight and the observation direction of the photometer: $\cos(\theta_{sca}) = \cos(\theta_{sza})\cos(\theta_{vza}) + \sin(\theta_{sza})\sin(\theta_{vza})\cos(\theta_{raa})$. For Almucantar diffused sky radiation observations parallel to the horizontal plane, there is only one solar zenith angle and one viewing zenith angle in one scan, and the two angles are equal.

Table 3. Output variables of the EML Model

Output variables	Count	Notes
Spectral SSA	4	Single scattering albedo of aerosols in four observation bands
Spectral g	4	Scattering asymmetric factor of aerosol in four observation bands
Effective radius r_{eff}	1	Characterize the particle size of the aerosol group in the atmosphere column
Fine mode fraction FMF	1	Characterization of the volume proportion of fine particles (with a radius less than 1 micron) in the aerosol group in the atmospheric column

121

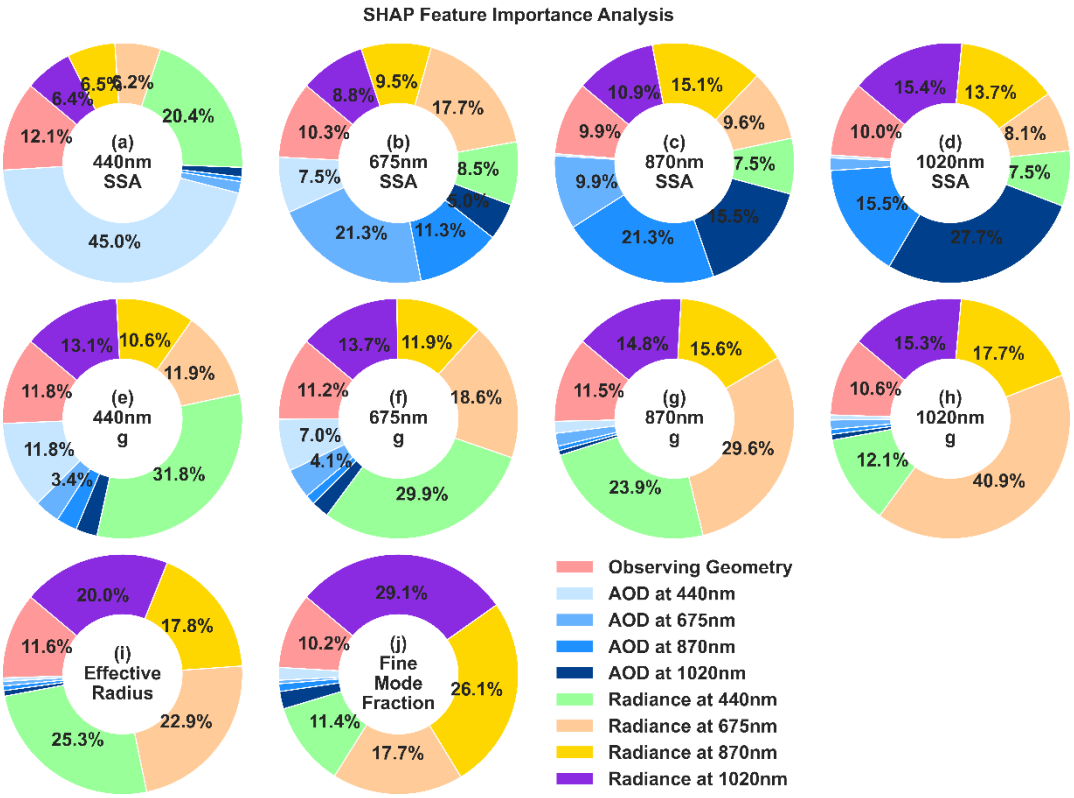

Figure 4. Importance analysis of input features based on SHAP values. Subfigures a-d correspond to retrieved variables SSA, e-h correspond to retrieved variables g, i correspond to r_{eff}, and j correspond to FMF. The four columns represent the observation wavelengths of 440, 675, 870, and 1020 nm in the first two rows. All 120 input features of the EML model are grouped into categories. Observation geometry includes the cosine of SZA and the scattering angle from the Almucantar scanning mode. Radiance refers to measured sky radiances from 23 observation geometries. Values less than 3% are hidden.

Reviewer Comment #5:

331, 407, 411 SCAs has this acronym defined earlier? I cannot find.

Response:

Sorry, it was indeed an omission in article writing process. SCA refers to "*the scattering angle*" and has been defined in the manuscript.

Reviewer Comment #6:

382-394 What is the reason to have lower performance for coarse particles as compared to fine particles, other than non-sphericity? Less number of training dataset? Please elaborate a little bit more.

Response:

We agree that non-sphericity alone may not fully explain the relatively lower retrieval performance for coarse particles, and we have expanded the discussion accordingly in the revised manuscript.

First, coarse-mode aerosols, such as dust and sea salt, are often significantly non-spherical. Radiative transfer calculations in many operational and machine-learning-based retrieval frameworks are typically based on Mie theory or spherical assumptions, which cannot fully reproduce the angular scattering patterns of irregular particles. As shown in previous studies (Dubovik et al., 2006), particle non-sphericity can substantially modify phase functions and polarization signals, especially in the coarse mode. In forward radiative transfer modeling, the T-matrix method—while capable of handling non-spherical particles—still faces limitations when it comes to accurately representing the scattering phase matrix of multi-faceted, multi-angle particles, which are approximated as ellipsoids (Mishchenko et al., 1996). This increases forward-model errors and reduces the uniqueness of the inverse solution, thereby degrading retrieval performance. In addition to non-sphericity, several other factors may contribute to the lower performance for coarse particles. (1) The radiative sensitivity of Sun–sky photometer measurements to coarse-mode microphysical variations is generally weaker than for fine-mode particles, particularly in forward scattering geometries, which reduces the information content available for inversion. Non-spherical scattering has higher asymmetry, far beyond forward and backward asymmetry, while ground-based photometers receive more forward scattering signals. (2) Strong parameter coupling among coarse-mode effective radius, volume concentration, and asymmetry factor may increase the ill-posedness of the inverse problem. (3) The distribution of training samples may also play a role, as coarse-mode-dominated cases are typically less frequent than fine-mode-dominated cases in ground-based observational datasets, potentially limiting the representation of extreme coarse regimes in the training process. We have now clarified these aspects in the revised manuscript to provide a more comprehensive interpretation of the observed performance differences.

The discussion added to the manuscript is as follows: *In addition, strong parameter coupling among coarse-mode effective radius, volume concentration, and asymmetry factor may increase the ill-posedness of the inverse problem. The distribution of training samples may also play a role, as coarse-mode-dominated cases are typically less frequent than fine-mode-dominated cases in observational datasets, potentially limiting the representation of extreme coarse regimes in the training process.*

Finally, thank you again for providing all the review comments, which helped me further consider and improve the algorithm design and manuscript content.

Reference:

Dubovik, O., Sinyuk, A., Lapyonok, T., Holben, B. N., Mishchenko, M. I., Yang, P., Eck, T. F., Volten, H., Muñoz, O., Veihelmann, B., van der Zande, V. J., Leon, J.-F., Sorokin, M., and Slutsker, I.: The application of spheroid models to account for aerosol particle nonsphericity in remote sensing of desert dust, J. Geophys. Res.-Atmos., 111, D11208, https://doi.org/10.1029/2005JD006619, 2006.

Mishchenko, M. I., Travis, L. D., Mackowski, D. W.: T-Matrix Computations of Light Scattering by Non-spherical Particles: A Review, J. Quant. Spectrosc. Radiat. Transfer, 55, 535–575, https://doi.org/10.1016/0022-4073(96)00002-7 , 1996.