https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

An Ensemble Machine Learning Method to Retrieve Aerosol Parameters from Ground-based Sun-sky Photometer Measurements

Qiurui Li[1], Zhongxia Sun[1,3], Meijing Liu[1], Huizheng Che[2], Yu Zheng[2], Jing Li[1*]

[1]Department of Atmospheric and Oceanic Sciences, Peking University, Beijing, 100871, China
[2]Key Laboratory of Atmospheric Chemistry, Chinese Academy of Meteorological Sciences, Beijing, 100081, China
[3]Paul Scherrer Institut, Forschungsstrasse 111, 5232 Villigen PSI, Switzerland

Correspondence to: Jing Li (jing-li@pku.edu.cn)

Abstract. Ground-based Sun-sky photometers have been widely used to measure aerosol optical and microphysical properties, yet the conventional numerical inversion schemes are often computationally expensive. In this study, we developed an explainable Ensemble Machine Learning (EML) model that simultaneously retrieves aerosol single scattering albedo (SSA), scattering asymmetry parameter (g), effective radius (r_{eff}), and fine-mode fraction (FMF) from direct and diffuse solar radiation measurements, with feature importance quantified using SHapley Additive exPlanations (SHAP). The EML model was trained and validated on a dataset of 110,000 samples simulated using the T-matrix particle scattering model and the VLIDORT radiative transfer model, encompassing diverse aerosol, atmospheric, and surface conditions. The algorithm demonstrated robustness through ten-fold cross validation, achieving correlation coefficients of 0.94, 0.95, 0.92, and 0.90 for SSA, g, r_{eff}, and FMF on the validation set, respectively. SHAP-based feature importance analysis confirmed the physical interpretability of the model, highlighting its effective use of multi-band radiance information and the stronger dependence of SSA retrieval on aerosol optical depth (AOD) relative to g and r_{eff}. Retrieval uncertainties estimated from repeated noise perturbation experiments were 0.03 for SSA, 0.02 for g, 0.08 for r_{eff}, and 0.09 for FMF. Applied to 132,067 sets of raw photometer measurements, the EML-based retrieval produced forward radiance fitting residuals comparable to those of the AERONET official inversion products. Moreover, compared with numerical algorithms, the EML model eliminates the need for a priori assumptions and smoothness constraints, while improving computational efficiency by more than five orders of magnitude.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

1 Introduction

Ground-based Sun-sky photometers are widely used remote sensing instruments for observing column-averaged aerosol optical and microphysical properties. The system typically measures direct solar irradiance, diffuse sky radiance, and the degree of linear polarization across multiple atmospheric window channels, spanning a broad range of scattering angles. They enable retrievals of aerosol optical depth (AOD), single scattering albedo (SSA), and particle size distribution, which are critical for characterizing aerosol loading, type, and radiative effects. The AErosol RObotic NETwork (AERONET, Holben et al., 1998) is the most successful global photometer network, operated by the National Aeronautics and Space Administration (NASA). Each AERONET site is equipped with a Cimel Electronique CE-318 photometer, which operates in three primary sky-scanning modes: Almucantar, Principal Plane, and Hybrid. In the Almucantar scan, the viewing zenith angle (VZA) is set equal to the solar zenith angle (SZA), whereas in the Principal Plane scan, the viewing azimuth angle is fixed to the solar azimuth angle. The Hybrid scan combines both approaches, beginning with Almucantar and then switching to Principal Plane scanning, thereby ensuring adequate scattering angle coverage even when SZA exceeds 50 °. Since its establishment in the early 1990s, AERONET has provided long-term, high-quality aerosol observations that have been extensively used for satellite data validation (Chu et al., 2002; Kahn et al., 2005; Levy et al., 2010; Omar et al., 2013; Fan et al., 2023), air quality monitoring (Dubovik et al., 2002; van Donkelaar et al., 2010; El-Nadry et al., 2019), and aerosol climate forcing studies (Garc á et al., 2012; Mao et al., 2019; Logothetis et al., 2021), among other applications.

AERONET has a standardized official inversion algorithm that utilizes Almucantar radiance observations at four wavelengths (440, 675, 870, and 1020 nm) to derive aerosol optical and microphysical parameters, including SSA, scattering asymmetry parameter (g), and effective radius (r_{eff}), among others. 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). SSA, g, and other aerosol optical parameters are subsequently calculated from the retrieved microphysical properties using Mie theory for spherical particles and the T-matrix approach for non-spherical particles (Dubovik et al., 2006). Similar networks have been established worldwide, providing complementary and more detailed information on regional aerosol characteristics. Examples include SKYNET in Asia and Europe (Takamura et al., 2004; Nakajima et al., 2003), the AERosol CANada（AEROCAN) in Canada (Bokoye et al., 2001), the Aerosol Ground Station Network (AGSNet) in Australia (Mitchell and Forgan, 2003), and the China Aerosol Remote Sensing Network (CARSNET) in China (Che et al., 2008, 2015). The main instrument of SKYNET is a sky radiometer, with observation wavelengths and scanning geometries similar to those of Sun–sky photometers. SKYNET aerosol retrievals are performed using the Skyrad Pack, which follows an inversion philosophy similar to that of the official AERONET algorithm. AEROCAN, AGSNet, and CARSNET employ the same Cimel photometers and inversion algorithms as AERONET. While the AERONET-type inversion algorithm achieves relatively high accuracy, it suffers from the need for a priori assumptions and limited computational efficiency. Retrieving aerosol size distribution from diffuse sky radiance is an ill-posed inverse problem: solutions are non-unique and unstable with respect to measurement noise. To regularize the inversion, the

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

algorithm imposes a priori assumptions and smoothness constraints, which suppress unphysical oscillations in the spectral dependence of the retrieved parameters (Dubovik and King, 2000). However, the choice of these constraints and their strengths is partly subjective and can introduce artificial biases. Furthermore, the computational cost of the numerical algorithm depends strongly on the initial guess and noise level. When the initial state is far from the truth and/or the observations are noisy, the inversion requires more radiative transfer calculations to reach convergence, thereby consuming significantly more time and, in some cases, even failing to converge. Previous improvements to the AERONET-type algorithm have mainly targeted forward radiative transfer calculations, including transitioning RTMs from scalar to polarized formulations, updating solar flux spectra and gas absorption databases, and accounting for non-spherical aerosols. However, these efforts cannot fully address the inherent limitation of low computational efficiency in numerical inversion algorithms (Sinyuk et al., 2020). Recently, rapid advances in machine learning have offered promising alternatives for remote sensing of atmospheric composition. Machine learning methods not only capture nonlinear relationships more effectively and operate far faster than numerical approaches, but also eliminate the need for initial guesses and prior constraints.

In the past few years, the field of aerosol remote sensing also experienced a bloom in machine learning algorithms. For satellite-based aerosol retrieval, machine learning approaches can be broadly divided into two categories according to the source of the training data: (1) those that pair satellite observations with AERONET aerosol products (Vucetic et al., 2008; Liang et al., 2020; Chen et al., 2022; Cao et al., 2023; Dong et al., 2024; She et al., 2024;), and (2) those that rely on RTM simulations tailored to the measurement configurations of satellite sensors (Sun et al., 2020; Qi et al., 2022; Tao et al., 2023). The first approach benefits from training data that closely represent real atmospheric conditions but is constrained by limited data volume and site representativeness. The second approach enables coverage of diverse atmospheric and aerosol types and supports the generation of large training datasets; however, models trained solely on simulations often face a substantial domain gap when applied to real observations, leading to a sharp performance drop. By comparison, only a few ML algorithms have been developed for ground-based aerosol retrieval, and most existing efforts use AERONET products as truth for training. For example, Cazorla et al. (2009) trained a neural network with AERONET AOD as reference to retrieve AOD from All-Sky Imager measurements. Huttunen et al. (2016) applied four machine learning models to estimate AOD from CM21 pyranometer measurements, but their validation against AERONET data was limited to the Thessaloniki site in Greece. Taylor et al. (2014) employed multi-band AOD, water vapor, and absorption AOD as inputs to a neural network to infer daily aerosol complex refractive index, SSA, and size distribution, thereby extending the scope of satellite remote sensing products. However, they did not use satellite or ground-based radiation measurements.

To date, no machine learning approach has been widely adopted for ground-based Sun-sky photometer inversions. This study develops an Ensemble Machine Learning (EML)-based aerosol retrieval algorithm that simultaneously retrieves SSA, g, r_{eff}, and fine-mode fraction (FMF) from CE-318 photometer measurements. We employ SHapley Additive exPlanations (SHAP) to quantify feature importance and provide physical insights into the retrieval process (Hou et al., 2022; Zhang et al., 2024). Instead of relying on co-located instrument measurements and products derived from existing algorithms, the training set is generated through forward radiative transfer simulations. The remainder of this paper is organized as follows. Sect. 2 describes

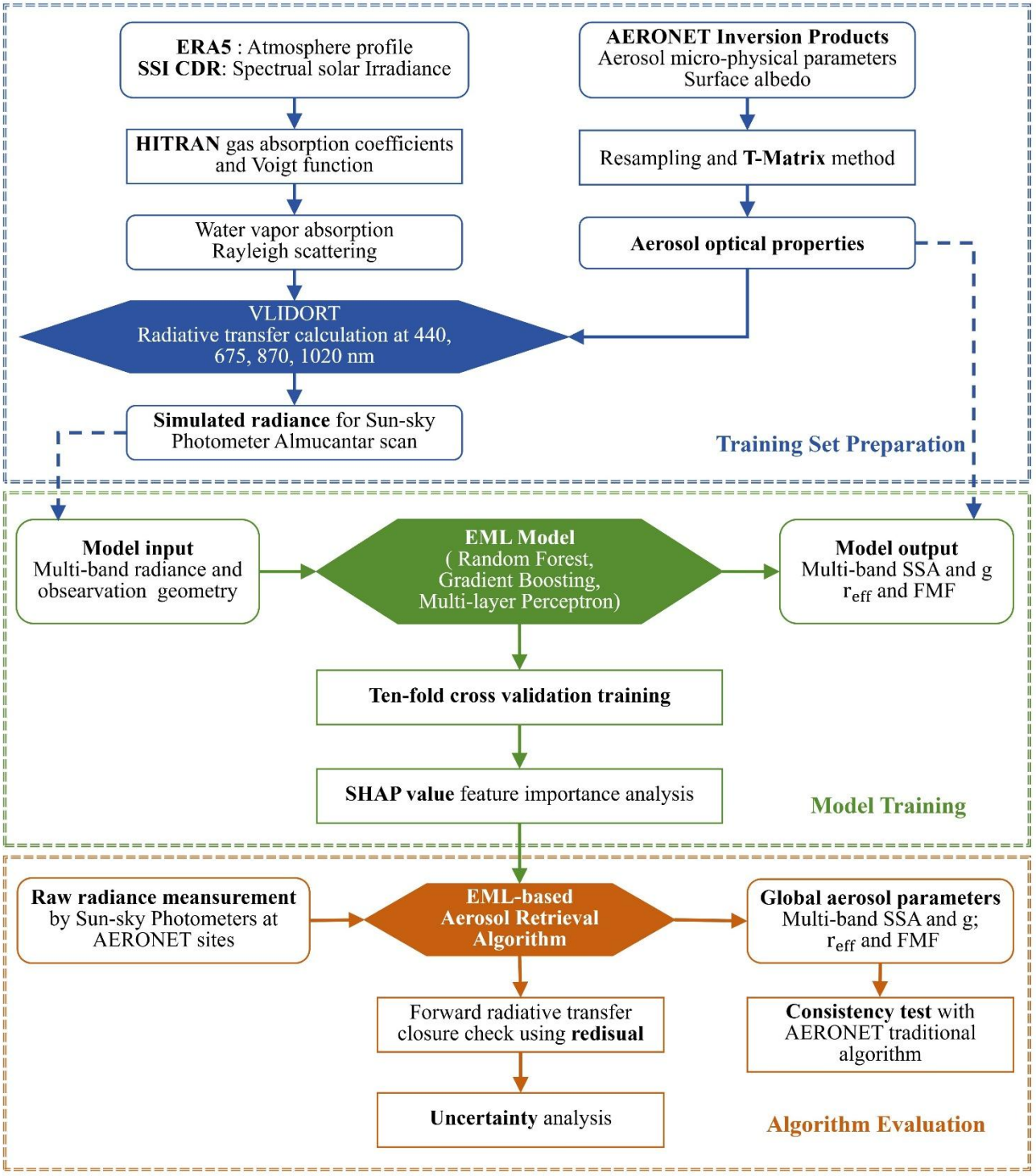
https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

the architecture of the proposed EML-based aerosol retrieval algorithm and the construction of the training, validation, and test datasets. Sect. 3 presents the results, including model fitting on simulated data, retrievals from raw measurements, SHAP-based feature importance analysis, and uncertainty evaluation. Finally, Sect. 4 summarizes the key features of the algorithm and discusses its advantages and potential applications in future aerosol remote sensing.

2 Data and Algorithm

Our proposed EML-based aerosol inversion algorithm is designed for the ground-based CE-318 Sun-sky photometer. The algorithm performs a joint retrieval at four observational wavelengths (440, 675, 870, and 1020 nm), simultaneously deriving SSA, g, r_{eff}, and FMF. It requires three types of inputs: (1) spectral AODs, (2) diffuse sky radiances from Almucantar scans at four wavelengths, and (3) geometric observation parameters, including SZA, VZA, and relative azimuth angle (RAA). An overview of the retrieval framework is shown in Fig. 1. The model is trained and validated on a large synthetic dataset generated through forward radiative transfer simulations, ensuring sufficient sample size and diversity. Independent testing is performed using photometer observations from AERONET sites, enabling assessment of both retrieval accuracy on real measurements and consistency with the official AERONET algorithm. In the following subsections, we describe (1) AERONET AOD and diffuse sky radiance measurements along with the associated inversion products, (2) the setup of forward radiative transfer simulations, (3) the design and implementation of the EML-based algorithm and the SHAP analysis, and (4) the methodology for estimating retrieval errors and uncertainties.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025



https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Figure 1. Flowchart of the EML-based aerosol retrieval algorithm for ground-based Sun-sky photometers. Colored oblong diamonds indicate models or algorithms, round-cornered rectangles represent input/output data, and regular rectangles denote processing steps.

2.1 AERONET Photometer Measurements and Aerosol Inversion Products

The ground-based Sun-sky photometer measures both direct and diffuse solar radiation. Direct solar irradiance is observed across ultraviolet, visible, and near-infrared bands, and AOD is retrieved from these measurements using the Beer–Lambert law after accounting for Rayleigh scattering and gaseous absorption. During Almucantar scans, diffuse sky radiance is recorded at 30 RAAs (2 °, 2.5 °, 3 °, 3.5 °, 4 °, 5 °, 6 °, 7 °, 8 °, 10 °, 12 °, 14 °, 16 °, 18 °, 20 °, 25 °, 30 °, 35 °, 40 °, 45 °, 50 °, 60 °, 70 °, 80 °, 90 °, 100 °, 120 °, 140 °, 160 °,180 °). AOD and radiance measurements at RAA greater than 7 °are used to retrieve aerosol parameters including SSA, *g*, size distribution, and refractive index (Dubovik and King, 2000). AERONET inversion products are classified into Level 1.0 (unscreened), Level 1.5 (cloud-screened and quality-controlled), and Level 2.0 (quality-assured). Level 2.0 data are produced through uniform instrument calibration and rigorous manual inspection, with quality control criteria such as AOD > 0.4, SZA > 50 °, and sky residual < 5%, which considerably reduces data volume but ensures high reliability. The uncertainties of Level 2.0 retrievals are typically about 0.03 for SSA and 0.02 for *g* (Giles et al., 2019; Sinyuk et al., 2020).

We downloaded coincident Level 2.0 AOD and aerosol inversion products, along with the corresponding raw Almucantar radiance measurements, from AERONET global sites to construct a testing set of 132,067 cases. This dataset was used to evaluate the retrieval capability of the proposed EML-based algorithm on real observations. To supplement aerosol types under low-AOD conditions, Level 1.5 inversion products were also collected and matched with their corresponding radiance and AOD observations, yielding an additional 87,144 cases. Aerosol size distributions, refractive indices, and surface albedo from the Level 2.0 and Level 1.5 inversion products were resampled and randomly combined to generate aerosol inputs for the forward radiative transfer simulations (Sect. 2.2), ensuring both parameter validity and statistical consistency with observed aerosol properties. In addition, radiation measurements were analyzed to characterize observational noise, which was then added to the training and validation sets (Sect. 2.3).

2.2 Forward Radiative Transfer Simulation

We employed VLIDORT v2.8.1, a linearized vector radiative transfer model, to simulate Almucantar observations from the photometer (Sect. 2.1), thereby generating a comprehensive training and validation dataset. VLIDORT computes the full Stokes vector [*I, Q, U, V*] for any specified viewing geometry and optical depth (Spurr, 2006). Here, *I* denotes radiance intensity, while *Q* and *U* represent linear polarization components. The model solves the radiative transfer equation for multilayer multiple scattering, requiring inputs such as solar spectral irradiance (SSI) at the top of atmosphere, surface albedo, and atmospheric and aerosol profiles. Its accuracy and flexibility make it well suited for simulating radiative measurements under diverse aerosol and atmospheric conditions.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

SSI is obtained from the Solar Spectral Irradiance Climate Data Record, which provides the solar energy flux reaching the top of Earth's atmosphere for different wavelengths. Observations indicate that SSI variability under stable solar conditions is very small (less than 0.3% on daily to annual timescales), with an even smaller impact on ground-based measurements. Therefore, a fixed SSI was adopted, with values of 1824.85, 1487.16, 970.44, and 689.27 W/m^2 at 440, 675, 870, and 1020 nm, respectively. Surface reflectance is treated as a Lambertian boundary, since ground-based observations are dominated by downward solar radiation, with minimal contribution from surface reflection. In our algorithm, surface reflectance is neither an inverted nor an input variable. It is only used in radiative transfer simulations, with values sampled from AERONET inversion products (Sect. 2.1).

Radiative transfer is also controlled by both the column loading and vertical distribution of aerosols and gas molecules. The aerosol particle size distribution is assumed to follow a bimodal lognormal volume distribution:

$$\frac{dV}{dlnr} = \frac{C_{Vf}}{\sqrt{2\pi}ln\sigma_f}\exp\left(-\frac{(lnr-lnr_{vf})^2}{2ln^2\sigma_f}\right) + \frac{C_{Vc}}{\sqrt{2\pi}ln\sigma_c}\exp\left(-\frac{(lnr-lnr_{vc})^2}{2ln^2\sigma_c}\right) \quad (1)$$

where C_V, r_V and σ denote the volume concentration, volume mean radius and geometric standard deviation, respectively, and the subscripts *f* and *c* represent fine and coarse modes. Many studies have shown that the scattering properties of particles can be fully characterized using only their r_{eff} and effective standard deviation (Hansen and Travis, 1974; Davies, 1974; Whitby, 1978; Ott, 1990; Mishchenko et al., 2004). The effective radius r_{eff} and FMF are calculated as:

$$r_{eff} = \frac{\int_{r_{min}}^{r_{max}} r^3\frac{dN(r)}{dlnr}dlnr}{\int_{r_{min}}^{r_{max}} r^2\frac{dN(r)}{dlnr}dlnr} \quad (2)$$

$$\mathrm{FMF} = \frac{\sum_{r_{min}}^{1\mu m}\frac{dV}{dlnr}dlnr}{\sum_{r_{min}}^{r_{max}}\frac{dV}{dlnr}dlnr} \quad (3)$$

Many aerosol types, particularly dust, are non-spherical, which significantly affects their scattering properties. To account for this, we employed the randomly oriented rotating ellipsoid model, a simple extension of the spherical model characterized by an additional axis ratio parameter. The T-matrix algorithm (Mishchenko and Travis, 1994) computes SSA, the scattering phase matrix, and other optical properties for ensembles of ellipsoidal particles. In radiative transfer simulations, aerosol parameters are averaged over various shapes, making the exact geometry of individual particles less critical; the optical characteristics are primarily determined by the overall axis ratio distribution (Mugnai and Wiscombe, 1986; Bohren and Singham, 1991; Mishchenko et al., 1997). The ellipsoid axis ratios were sampled according to the probability distribution observed for typical dust events (Dubovik et al., 2006). The aerosol extinction coefficient, β, decays exponentially with height:

$$\beta(h) = \beta_0 e^{-h/H} \quad (4)$$

where h is the altitude and H is the extinction scale height, ranging from less than 1 km in winter to more than 2 km on turbid summer days (Turner et al., 2001). Atmospheric profile information was obtained from the ERA5 (European Centre for

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Medium-Range Weather Forecasts Reanalysis Version 5) monthly mean data (2020–2024) on pressure levels, including temperature, specific humidity, and ozone mass mixing ratio. Data from low- to mid-latitude land areas were extracted and spatially thinned to a $5° \times 5°$ grid to serve as the sampling database. Based on these meteorological fields, Rayleigh scattering and gas absorption were calculated. The Rayleigh scattering optical thickness τ_R at a specific visible wavelength λ was computed using the empirical formula of Dutton et al. (1994):

$$\tau_R(\lambda) = \frac{pressure}{1013.25\ hPa} \times 0.00877 \times \lambda^{-4.05} \quad (5)$$

which strictly applies under an exponentially decreasing atmospheric density. Water vapor and ozone absorption coefficient were calculated using the High-resolution Transmission Molecular Absorption Database (HITRAN). A Voigt line shape (Armstrong et al., 1967), accounting for both Doppler and pressure broadening, was applied to accurately model gas absorption under varying temperature and pressure conditions.

2.3 Inversion Architecture Using Ensemble Machine Learning

The EML has emerged as a powerful approach for capturing complex nonlinear relationships among variables by integrating multiple machine learning models, thereby leveraging their strengths while compensating for individual limitations. In this study, three base learners were adopted to construct the EML-based retrieval algorithm: Random Forest, Gradient Boosting, and Multi-Layer Perceptron. Random Forest represents a bagging approach that aggregates predictions from multiple decision trees trained on randomly sampled subsets of data and features (Breiman, 2001). 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). 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).

To enhance robustness, Gaussian white noise was injected into the training dataset. Proper noise perturbation is essential: too little noise reduces resistance to real-world observational errors, while too much can obscure true patterns. Noise characteristics were derived by comparing raw Almucantar observations with corresponding VLIDORT simulations based on AERONET inversion products (Sect. 2.1). From these differences, the signal-to-noise ratio was calculated to estimate the mean amplitude and standard deviation of the noise. Because solar radiation strongly depends on wavelength and angle, noise parameters vary with wavelength and RAA. Moreover, diffuse sky radiance spans a wide dynamic range, from about 10^{-1} W/m^2/sr at large angles to over 10^2 W/m^2/sr at small angles. To address this, all input and output variables were standardized to the interval $[-1, 1]$.

Ten-fold cross-validation (CV) was performed on the 100,000-sample training set to assess the EML model's generalization performance, with results summarized in Table 1 and discussed in Sect. 3.1. In this procedure, the training set is partitioned into ten equal subsets, and the model is iteratively trained on nine subsets while validated on the remaining one, repeating the

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

process until each subset has served as the validation set once. After CV, the final EML model was trained on the entire training set to fully leverage all available data.

To ensure physical interpretability, the EML-based inversion algorithm incorporates SHAP, a game-theoretic method that attributes model outputs to individual features while accounting for feature interactions (Zhao et al., 2019; Hou et al., 2022; Wang et al., 2023; Zhang et al., 2024). The SHAP value for a feature X_j is defined as:

$$\phi_j = \sum_{s \in N} \frac{|S|!(p-|S|-1)!}{p!} [f(S \cup \{j\}) - f(S)] \quad (6)$$

where p is the total number of features, N is the set of all feature subsets excluding X_j, S is a subset of N, $f(S)$ denotes the model prediction based on features in S, and $f(S \cup \{j\})$ is the prediction when X_j is added. The difference $[f(S \cup \{j\}) - f(S)]$ represents the marginal contribution of X_j for that subset, and the SHAP value ϕ_j is the weighted average of these contributions across all subsets. A larger SHAP value indicates a stronger influence of the feature on the model's predictions.

2.4 Model Evaluation and Uncertainty Estimation

Six statistical metrics were used to evaluate the predictive performance of the EML-based retrieval algorithm: correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), mean absolute deviation (MAD), linear bias, and error envelope (EE). These metrics quantify the agreement between the true values y and the predicted values $\hat{y}$:

$$R = \frac{\text{Covariance}(y,\hat{y})}{Variance(y)Variance(\hat{y})} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (11)$$

$$EE = \frac{\#\{\, y \mid |\hat{y} - y| < \pm\text{uncertainty} \,\}}{n} \quad (12)$$

where n is the number of cases. The uncertainty thresholds for EE follow the standards of existing ground-based aerosol inversion algorithms (Dubovik et al., 2000), with reference values of 0.03 for SSA, 0.02 for g, 0.1 for r_{eff}, and FMF.

The total inversion uncertainty σ was decomposed into systematic error σ_s and propagation error σ_p. Systematic error arises from the ill-posed nature of the inversion problem and the inherent limitations of the retrieval algorithm, and was quantified by applying the algorithm to the noise-free validation set, thereby excluding propagation effects. Propagation error results from the forward propagation of observational uncertainties and was evaluated through perturbation experiments. Gaussian

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

perturbations (100 realizations) were applied to the model input variables to simulate random observational errors, and the standard deviation of the resulting outputs was taken as σ_p. Perturbation magnitudes were scaled according to the uncertainty of each variable: geometric angles were assumed exact, AOD was assigned an absolute uncertainty of $\frac{1}{m}$ (where m is the optical air mass), and radiance was assumed accurate to within 5% across all wavelengths (Holben et al., 1998; Eck et al., 1999). The total uncertainty σ was then calculated as the quadratic mean of σ_s and σ_p:

$$\sigma = \sqrt{\sigma_s^2 + \sigma_p^2} \tag{13}$$

In theory, if the aerosol parameters retrieved by the algorithm are sufficiently accurate, they can be input into the RTM to reproduce the raw photometer measurements. The discrepancy between the simulated sky radiance y from the RTM and the observed radiance y^*, expressed in logarithmic scale, is defined as the optical residual:

$$\text{Residual (\%)} = \sqrt{\frac{\sum_{i=1}^{N}(lny^* - lny)^2}{N}} * 100 \tag{14}$$

where N denotes the total number of sky radiance observations in a single Almucantar scan. In this study, N=64, corresponding to radiance measurements at four wavelengths with RAAs greater than 20 °.

In addition, the relative deviation is defined as the difference between the observed radiance y^* and the simulated radiance y at a specific angle within a given band:

$$\text{Relative Deviation} = \frac{y^* - y}{y^*} * 100\% \tag{15}$$

This metric is used in Sect. 3.4 and illustrated in Fig. 7. Since the algorithm does not directly retrieve the complete aerosol size distribution required for radiative transfer calculations, the distribution was reconstructed using six-dimensional nearest-neighbor interpolation. The look-up table was generated from 110,000 sets of aerosol parameters prepared during the construction of the training and validation dataset. Its six search dimensions consist of g at four wavelengths, r_{eff}, and FMF.

3 Results

3.1 Model Fitting and Validation

The training and validation of our model are entirely based on the simulated dataset generated using the forward RTM. This design avoids dependence on instrument measurements or existing inversion products, and instead anchors the algorithm in radiative transfer theory for aerosol-laden atmospheres under clear-sky conditions. The performance of the EML model in the ten-fold CV is summarized in Table 1. The prediction scores remain highly consistent across folds, with variations within 0.01, which highlights the stability and robustness of the algorithm. This consistency further indicates that the algorithm maintains reliable predictive skill regardless of data partitioning. The average R^2, RMSE, and MAD are 0.773, 0.43, and 0.282, respectively. While the RMSE appears larger than the typical inversion uncertainties reported for individual aerosol parameters

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

(e.g., 0.03 for SSA and 0.02 for g), this is expected because these metrics aggregate deviations across all retrieved variables, rather than assessing each parameter independently.

Table1. Prediction Scores of EML Model via Ten-fold CV

Fold	1	2	3	4	5	6	7	8	9	10	Average
R^2	0.768	0.773	0.773	0.766	0.774	0.773	0.778	0.778	0.773	0.769	0.773
RMSE	0.433	0.428	0.426	0.435	0.429	0.434	0.433	0.426	0.425	0.429	0.430
MAD	0.284	0.281	0.280	0.284	0.283	0.282	0.280	0.283	0.227	0.227	0.282

The inversion performance on the validation set is presented in Fig. 2. As noted in Sect. 2.1, the validation dataset contains 10,000 independent cases generated by forward radiative transfer simulations, excluded from training but constructed with the same noise characteristics. The results confirm that the EML-based algorithm retrieves SSA, g, r_{eff}, and FMF simultaneously across four wavelengths with high accuracy and without evidence of overfitting. The scatter points are tightly distributed around the 1:1 line, indicating minimal systematic bias. Among the retrieved parameters, SSA achieves the strongest performance, with an *EE* of about 90%, an RMSE near 0.02, and *R* above 0.90. For SSA and g, the reported error statistics (e.g., RMSE) are wavelength-averaged. The asymmetry parameter g exhibits a slightly lower *EE* (~70%), which can be attributed to its stricter uncertainty threshold and increased bias at longer wavelengths. Nevertheless, g still achieves reasonable accuracy, with *R* around 0.95 and RMSE around 0.018. For the microphysical parameters r_{eff} and FMF, the *EE* values are approximately 75% and 66%, respectively, with both parameters showing *R* above 0.9. Overall, these results suggest that the algorithm achieves satisfactory retrieval performance across the validation set, with errors generally within acceptable bounds.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

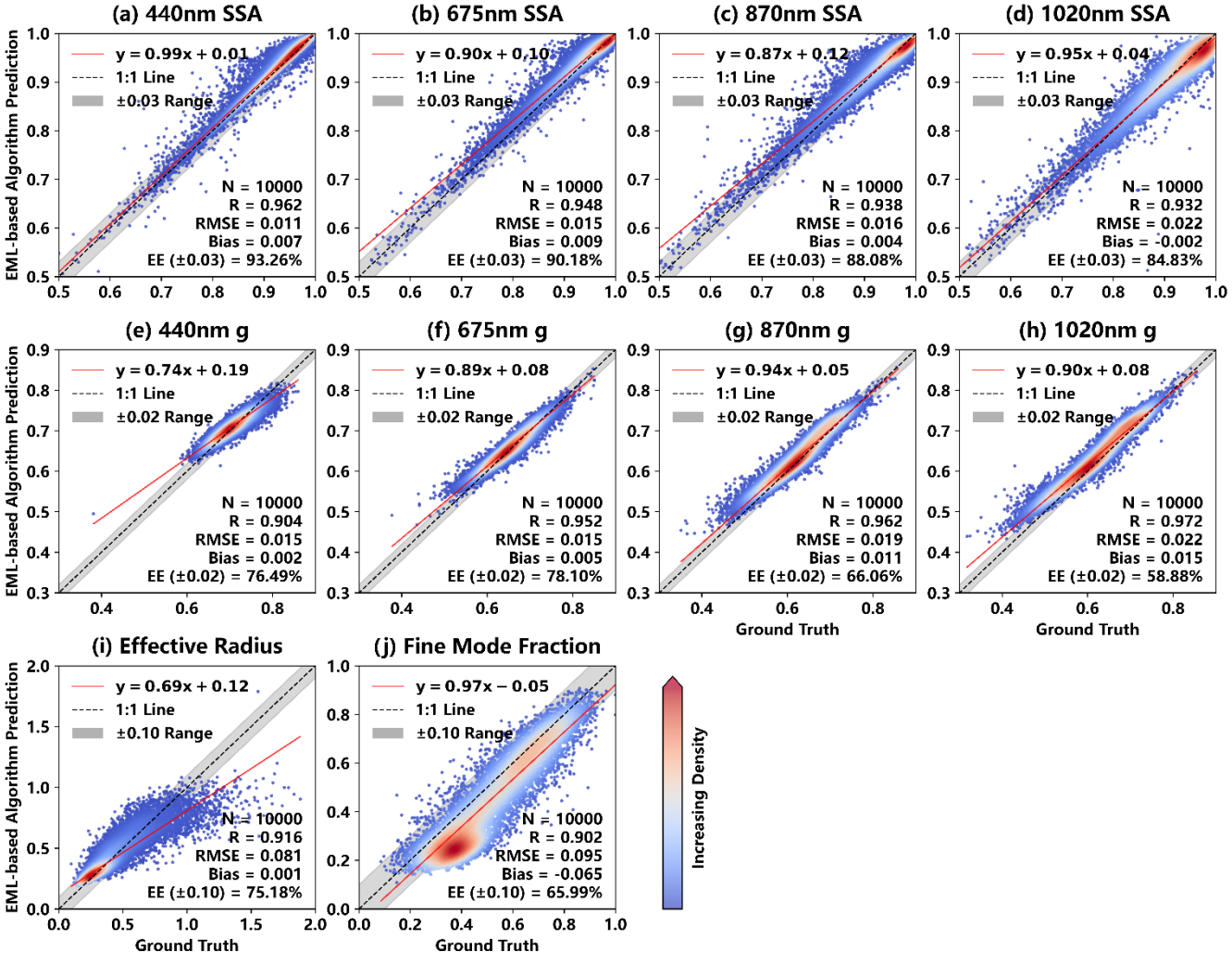


Figure 2. Aerosol parameters retrieved by the trained EML model versus the ground truth on the validation set. The color of the scatter points indicates point density. The four rows correspond to the four retrieved variables: SSA, g, r_{eff}, and FMF. The four columns represent the observation bands at 440, 675, 870, and 1020 nm. 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.

3.2 Retrieval Results on Raw Photometer Measurements

To further test the real-world applicability of our EML-based retrieval algorithm, we applied the model to ground-based photometer observations and compared the retrieved parameters with those from AERONET. This testing set comprises 132,067 cases derived from AERONET Level 2.0 inversion products paired with raw Almucantar sky radiance measurements, entirely excluded from model training and validation. Figure 3 shows the comparison results, with data points diluted by one-

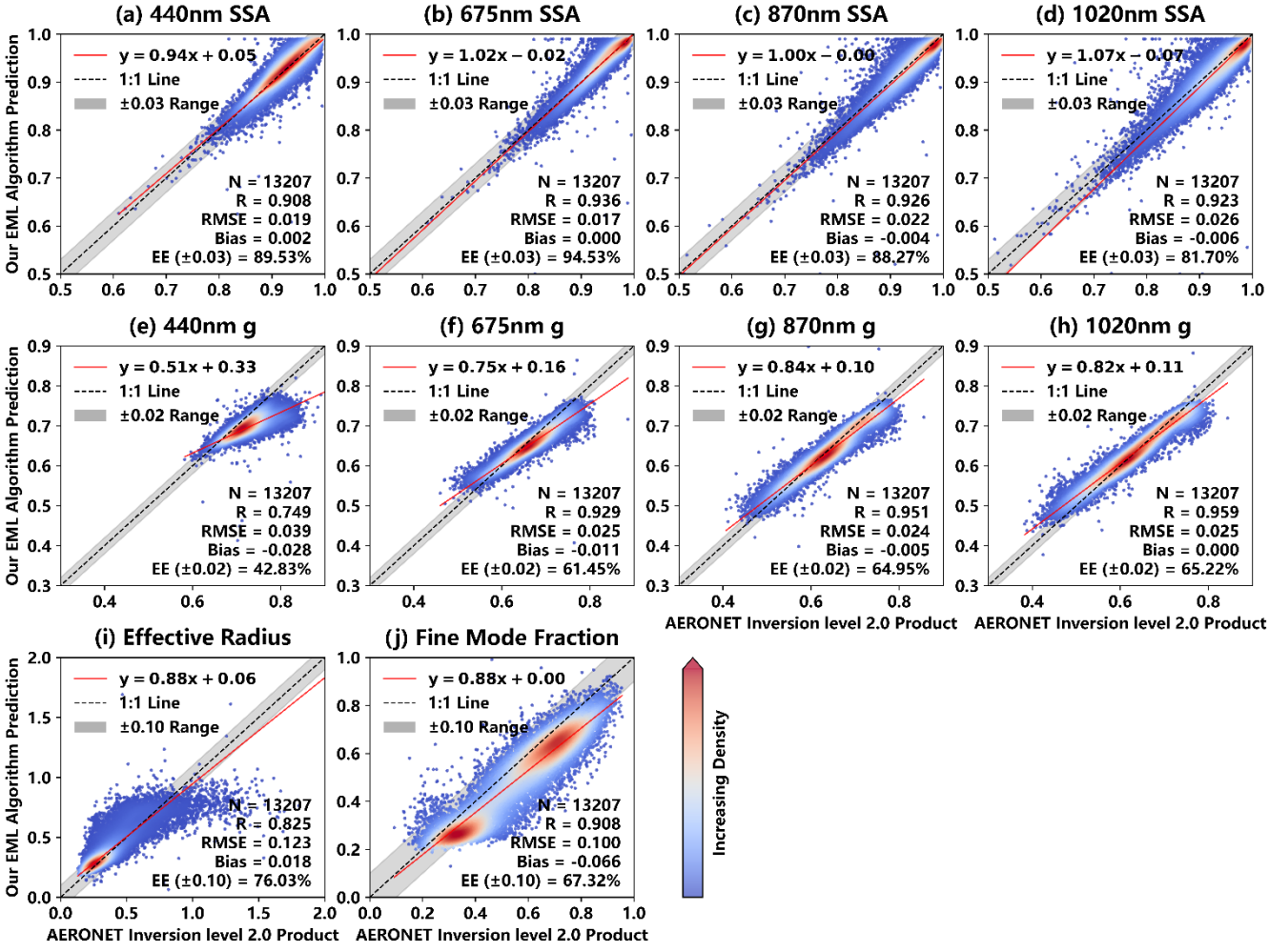
https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

tenth to improve visualization. The EML-retrieved parameters exhibit strong agreement with the AERONET products. Except for g at 440 nm, the R for all variables exceeds 0.9. The RMSEs of SSA and g are within 0.03, while those for r_{eff} and FMF are approximately 0.1. A notable advantage of the EML-based algorithm is its computational efficiency. It requires only 0.18 milliseconds to invert a single measurement, whereas traditional numerical retrieval algorithms often take several minutes per case. Dubovik et al. (2011) attempted to accelerate numerical inversion by optimizing forward radiative transfer calculations, such as reducing terms in the phase matrix expansion and quadrature integration. However, the time required for a complete retrieval still remained at the minute scale. In contrast, by eliminating iterative radiative transfer calculations, our algorithm increases the retrieval speed by a factor of ~10^5 compared with conventional numerical inversion schemes.

Regarding wavelength dependence, the retrieval accuracy for SSA decreases with increasing wavelength λ in both the validation set (Fig. 2) and the testing set (Fig. 3), whereas the accuracy for g improves. As λ increases, the aerosol size parameter ($x = \frac{2\pi r}{\lambda}$) decreases, leading to weaker single scattering and stronger multiple scattering in the total radiation field at longer wavelengths (Moosmüller et al., 2009; Moosmüller and Sorensen, 2018), which makes SSA more difficult to constrain. The relatively poorer performance of SSA retrieval at 440 nm observed in Fig. 3 may be attributed to the higher AOD uncertainty at this wavelength, which serves as input for both our EML-based algorithm and the AERONET official algorithm. Specifically, the AOD uncertainty is approximately ±0.01 for $\lambda > 440$ nm and ±0.02 for $\lambda \leq 440$ nm (Holben et al., 1998; Eck et al., 1999). The improved retrieval accuracy of g at longer wavelengths can be explained by two mechanisms. First, the sensitivity of the radiative transfer equation to g, as quantified by the magnitude or norm of the Jacobian matrix ($\frac{\partial I}{\partial g}$), increases with wavelength (Hasekamp and Landgraf, 2005; Kokhanovsky, 2013). At longer wavelengths, the range of retrieved g values broadens noticeably, as illustrated in Fig. 2 and Fig. 3. Second, the influence of aerosol size distribution on g becomes more pronounced at longer wavelengths. The forward-scattering peak of the phase function broadens with increasing λ, enhancing sensitivity to coarse-mode particles (Osborne et al., 2008; Kalashnikova and Sokolik, 2013). Consequently, retrieval errors for g decrease from about ±0.05 in the visible to ±0.02 in the near-infrared (Dubovik et al., 2006). This trend is also reflected in Fig. 3, where the RMSE of g decreases from 0.039 at 440 nm to 0.025 at 1020 nm.

Retrieving aerosol microphysical parameters is generally more challenging than deriving optical properties, and the retrieval accuracy of r_{eff} slightly decreases in the testing set relative to the validation set. Both r_{eff} and FMF are frequently recognized as key indicators of aerosol size distribution: fine-mode aerosols, such as sulfates, nitrates, and biomass burning particles, dominate when $r_{eff} < 0.3$ μm and FMF > 0.5, whereas coarse-mode aerosols, typically originating from natural sources like mineral dust and sea salt, prevail when $r_{eff} > 1.0$ μm and FMF < 0.3. In Fig. 3, FMF exhibits two distinct peaks near 0.3 and 0.7, corresponding to r_{eff} values of 0.6 and 0.28 μm, representing the coarse and fine modes, respectively. These results indicate that our algorithm can provide a basic classification of aerosols based on their retrieved optical properties (SSA and g) and size distribution (r_{eff} and FMF).

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025



315 **Figure 3. Aerosol parameters retrieved by the EML-based algorithm compared with AERONET Level 2.0 inversion products on the testing set.** The plot configuration is the same as in Fig. 2. 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.

3.3 Feature Importance Analysis

The normalized feature importance of input variables on the predicted outputs was quantitatively assessed using SHAP values, 320 as shown in Fig. 4. First, the EML model effectively extracts and utilizes band-specific observational data for aerosol parameter retrieval at the corresponding wavelengths, as evidenced by the fact that radiance at a given wavelength exhibits the highest SHAP value when inverting SSA or *g* at the same wavelength. For instance, the radiance at 440 nm shows the highest feature importance for retrieving SSA at 440 nm (20.3%), which is markedly greater than its contribution to SSA at other wavelengths. Similarly, when retrieving *g* at 440 nm, its feature importance reaches 31.8%, again clearly exceeding its importance for *g* at 325 other wavelengths. Second, the SHAP values for each retrieved parameter indicate that the EML model also leverages

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

observations across all wavelengths, particularly for g and r_{eff}, reflecting the physical relationship between aerosol properties, such as particle size, and the spectral dependence of scattered radiation. Third, when inverting SSA, AOD shows the highest feature importance, consistently exceeding 40%. This is expected because SSA is defined as the ratio of scattering to total extinction (scattering plus absorption), making accurate AOD essential for SSA retrieval from sky diffuse radiation measurements. In contrast, the importance of AOD diminishes when predicting r_{eff} and FMF, whereas sky diffuse radiance across multiple bands and SCAs becomes more influential. According to Mie scattering theory, scattering phase functions differ substantially between fine- and coarse-mode aerosols, which increases the sensitivity of measured scattered radiation to particle size. For ground-based observations, diffuse radiance predominantly arises from aerosol forward scattering and stronger diffuse radiance indicates greater forward-backward scattering asymmetry, suggesting a larger column-averaged aerosol radius. Finally, auxiliary observation geometry information (SZA, VZA, and RAA) also plays a critical role in retrieving all aerosol parameters. These variables control both the magnitude and angular distribution of the measured radiance, thereby directly affecting the radiative transfer pathlength and scattering regime characterization. Consequently, the importance associated with observation geometry remain stable at around 10% across all retrieval targets. Overall, the SHAP-based feature importance analysis demonstrates that the EML-based retrieval model successfully captures the underlying physical processes governing aerosol scattering of solar radiation, supporting its applicability for broader aerosol retrieval practices.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

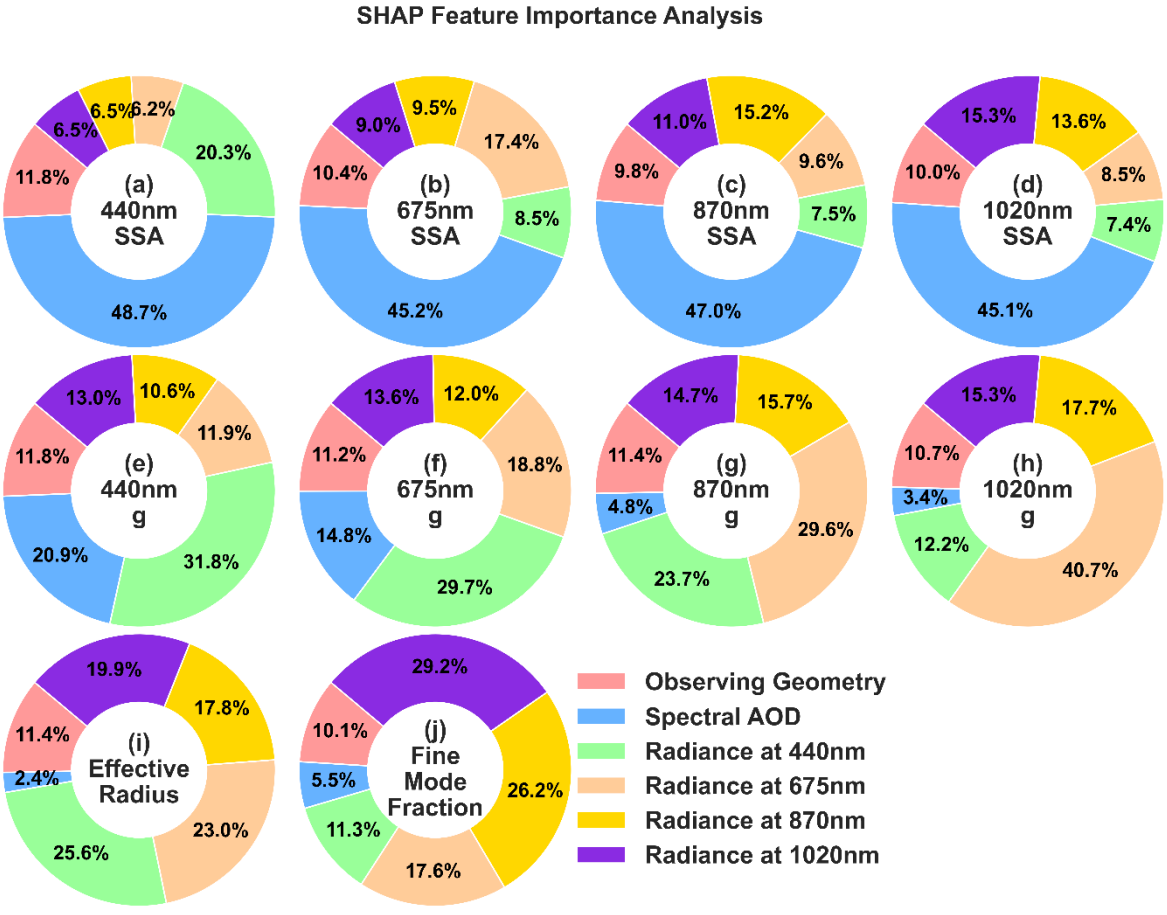


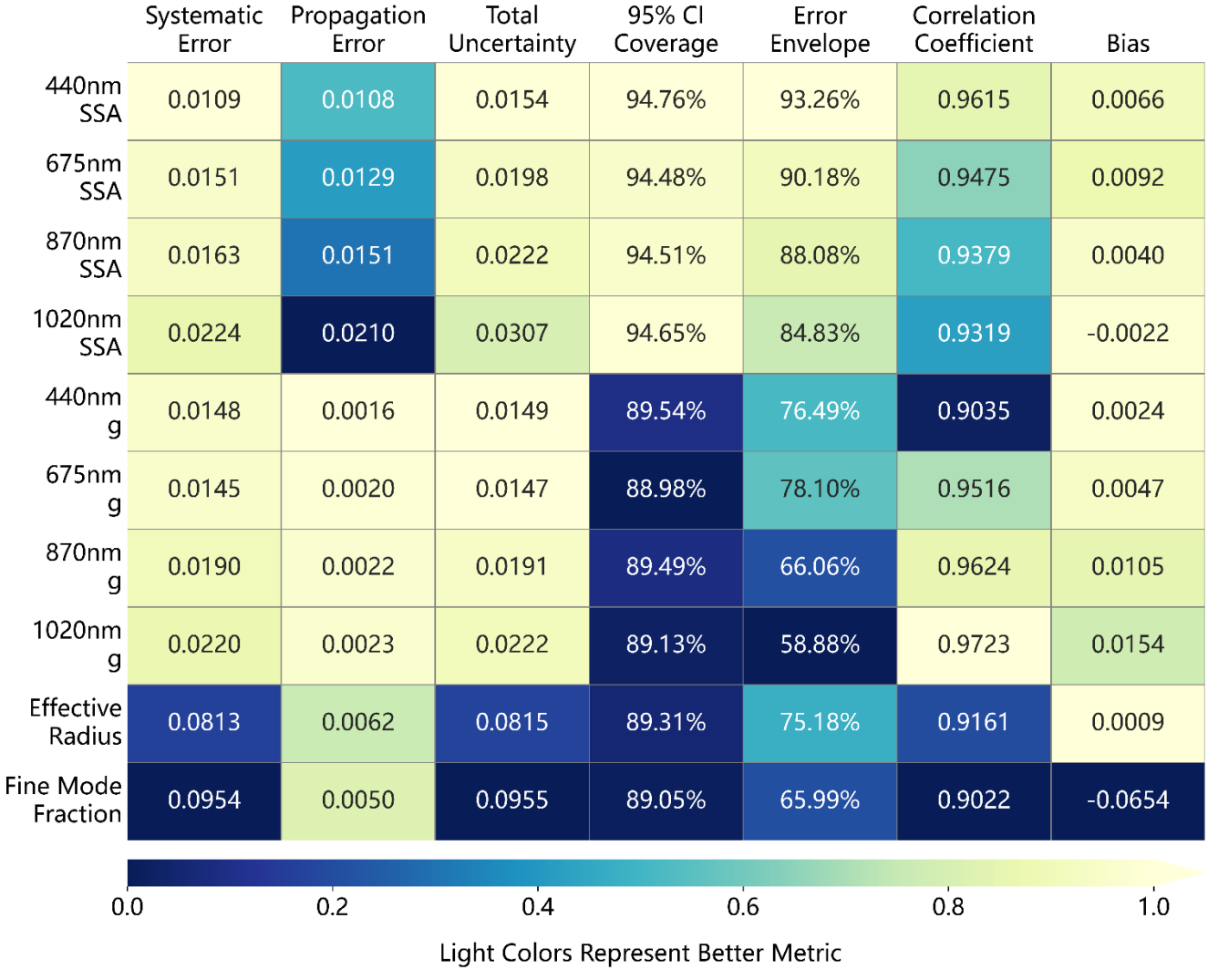
Figure 4. Importance analysis of input features based on SHAP values. The four rows correspond to the four retrieved aerosol parameters: SSA, g, r_{eff}, and FMF. The four columns represent the observation wavelengths of 440, 675, 870, and 1020 nm. All 120 input features of the EML model are grouped into categories. Observation geometry includes the cosine of SZA and the scattering angle from the Almucantar scanning mode. AOD denotes the aerosol optical depth at the four wavelengths. Radiance refers to measured sky radiances from 23 observation geometries.

3.4 Error evaluation and Uncertainty Analysis

We quantify the uncertainties in retrieving SSA, g, r_{eff}, and FMF with the EML-based aerosol retrieval algorithm using the method described in Sect. 2.4. Systematic errors are defined as the RMSE of retrievals from the noiseless validation set, whereas propagation errors are estimated from the standard deviation of retrieval variability across 100 noise-perturbed realizations of AOD and radiance. As shown in Fig. 5, the two types of errors are comparable in magnitude for SSA, while for the other parameters the systematic errors exceed the corresponding propagation errors. The total absolute uncertainties for SSA and g both tend to increase with wavelength. Specifically, for SSA the uncertainties are 0.0154, 0.0198, 0.0222, and

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

0.0307 at 440, 675, 870, and 1020 nm, respectively, while for g they are 0.0149, 0.0147, 0.0191, and 0.0222 at the same wavelengths. For the microphysical parameters, the total uncertainties are 0.082 for r_{eff} and 0.096 for FMF. These levels are comparable to those reported for existing aerosol inversion algorithms. For example, the official AERONET algorithm reports uncertainties of 0.02–0.03 for SSA and about 0.02 for g (Dubovik et al., 2002), while relative uncertainties in r_{eff} can exceed 20% due to the complexity of aerosol mixing states (Andrews et al., 2017). The 95% confidence interval (CI) coverage measures the probability that the true parameter value lies within the model-predicted uncertainty range for a single noise-perturbed inversion case, whereas the *EE* denotes the fraction of cases that satisfy the predefined uncertainty criteria. Both metrics decrease in the order SSA > g > r_{eff} > FMF, indicating that, compared to aerosol optical parameters, the retrieval of microphysical parameters generally requires higher observation data quality and greater algorithmic accuracy.



https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Figure 5. Heatmap of aerosol inversion uncertainties using the EML-based retrieval algorithm. The color scale does not represent absolute values; instead, lighter colors indicate better model performance for the corresponding variable based on the indicators in each column. The correlation coefficient and bias values are directly taken from Fig. 2.

We further evaluated the capability of our EML-based retrieval algorithm by using the aerosol parameters it retrieves to reproduce photometer observations. The accuracy of these retrieved parameters is reflected in the optical residual, which quantifies the discrepancy between the RTM-simulated radiance and the observed photometer measurements (see Sect. 2.4 for the detailed definition). Smaller optical residuals indicate higher retrieval accuracy, providing a quantitative measure of the retrieval quality. This assessment was performed using the testing set described in Sect. 2.1. Site-averaged retrieval residuals from our algorithm were compared with those from the AERONET official algorithm in Fig. 6. Across most sites, the residual magnitudes of the two algorithms are consistent, with differences generally within ±4% (Fig. 6c). From the perspective of algorithm design, the AERONET-type numerical algorithm minimizes the optical residual as a convergence criterion, whereas the EML model is trained to minimize the RMSE between predicted aerosol parameters and their reference values. That the EML-based algorithm achieves residual magnitudes comparable to the physics-based AERONET algorithm underscores its reliability.

Spatially, both algorithms exhibit similar residual distribution patterns: smaller residuals are observed over North and South America, East Asia, and Europe, whereas larger residuals occur over dust source regions such as North Africa and the Arabian Peninsula. Interestingly, the spatial pattern of residual differences between the two algorithms mirrors that of the mean r_{eff} retrieved by the EML model. Notably, the spatial pattern of residual differences between the two algorithms closely resembles that of the mean r_{eff} retrieved by the EML model, highlighting that the model's performance is less certain in regions dominated by coarse, non-spherical particles and pointing to potential areas for improvement. Sites in North Africa, South Asia, and inland China—where coarse-mode aerosols such as dust prevail—exhibit higher retrieval uncertainties. This effect is most pronounced at the shortest wavelength (440 nm, Fig. A1), where aerosol scattering exerts the strongest influence. Although both algorithms account for non-spherical particle scattering, neither fully resolves this complexity, indicating that further algorithmic refinement is needed. Additionally, some stations display substantially higher residuals relative to neighboring sites. At these locations, observational data are often sparse, potentially due to limited instrument maintenance or calibration. In certain cases, such as at some European sites, consistently low aerosol loading means the AOD rarely exceeds the 0.4 threshold required for AERONET Level 2.0 inversion products, contributing to larger residuals (Fig. B3).

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

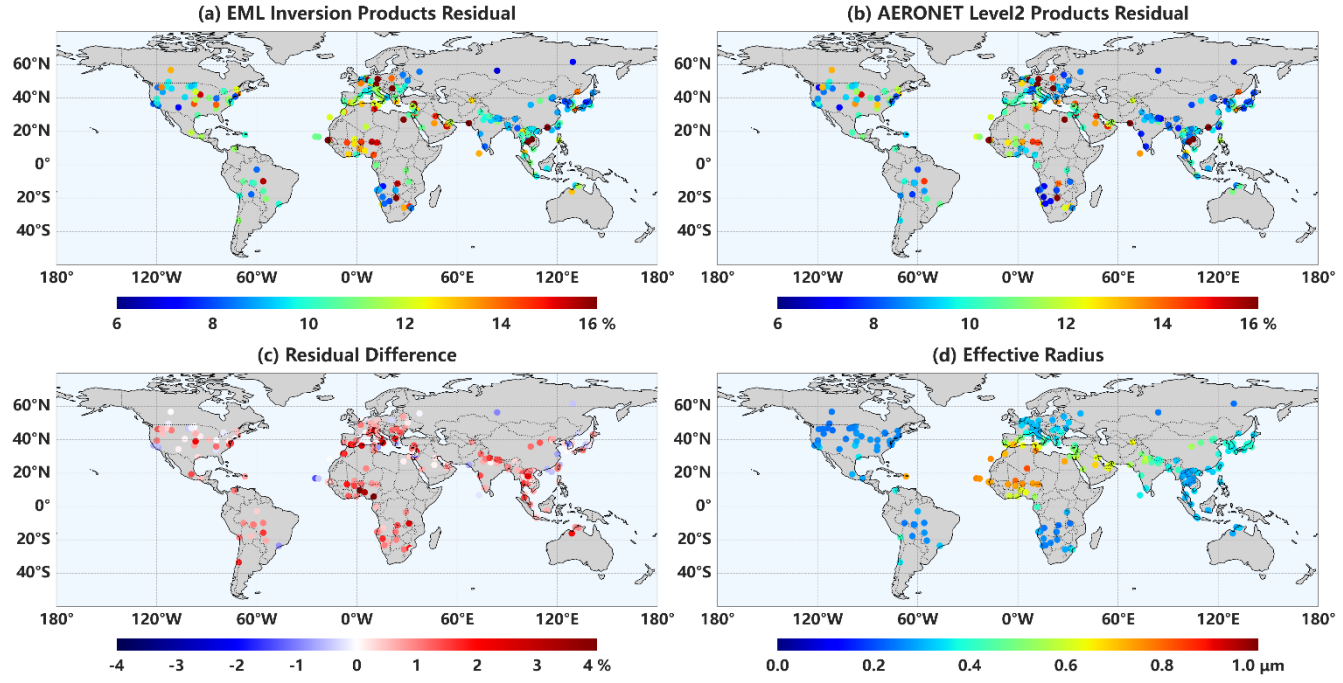


Figure 6. Site-averaged optical residuals for our EML-based and AERONET official aerosol inversion algorithms on the testing set. The residuals for all cases at each site were averaged, and the difference is calculated as the EML inversion product residual minus the AERONET level 2.0 product residual. The r_{eff} values were retrieved using the EML-based aerosol retrieval model developed in this study and subsequently averaged at each site.

Figure 7 shows the relative deviation between radiances simulated from the inversion results and those observed by the photometer, plotted as a function of RAA. Across all four observation wavelengths, the relative deviation exhibits a similar dependence on RAA. Minimal deviations (< 10%) and peak correlation coefficients (> 0.95) are observed at RAAs between 20 °and 100 °, indicating optimal agreement within this angular range. The current AERONET V3 retrieval algorithm excludes measurements with RAA < 20 °to minimize cloud contamination and forward-scattering effects (Giles et al., 2019). Similarly, the SKYNET algorithm prioritizes radiance observations within SCAs of 20 °–70 °for aerosol property retrieval (Nakajima et al., 1996, 2020). For a SZA of 60 °, RAAs between 20 °and 100 °correspond to SCAs of approximately 17 °–83 °. These SCA ranges align closely with those designed for passive visible-light remote sensing sensors, such as MODIS (Levy et al., 2013), VIIRS (Hsu et al., 2019), and POLDER (Deschamps et al., 1994).

Physically, a broader SCA range generally provides more information for the inversion of aerosol optical and microphysical properties. However, very small RAAs increase the likelihood of interference from direct solar radiation, and Sun-sky photometer measurements with RAA < 7 °are often overexposed or saturated. Conversely, as RAA approaches 180 °, the

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

photon flux along single-scattering paths diminishes, leading to a sharp drop in the measured radiance and a lower signal-to-noise ratio.

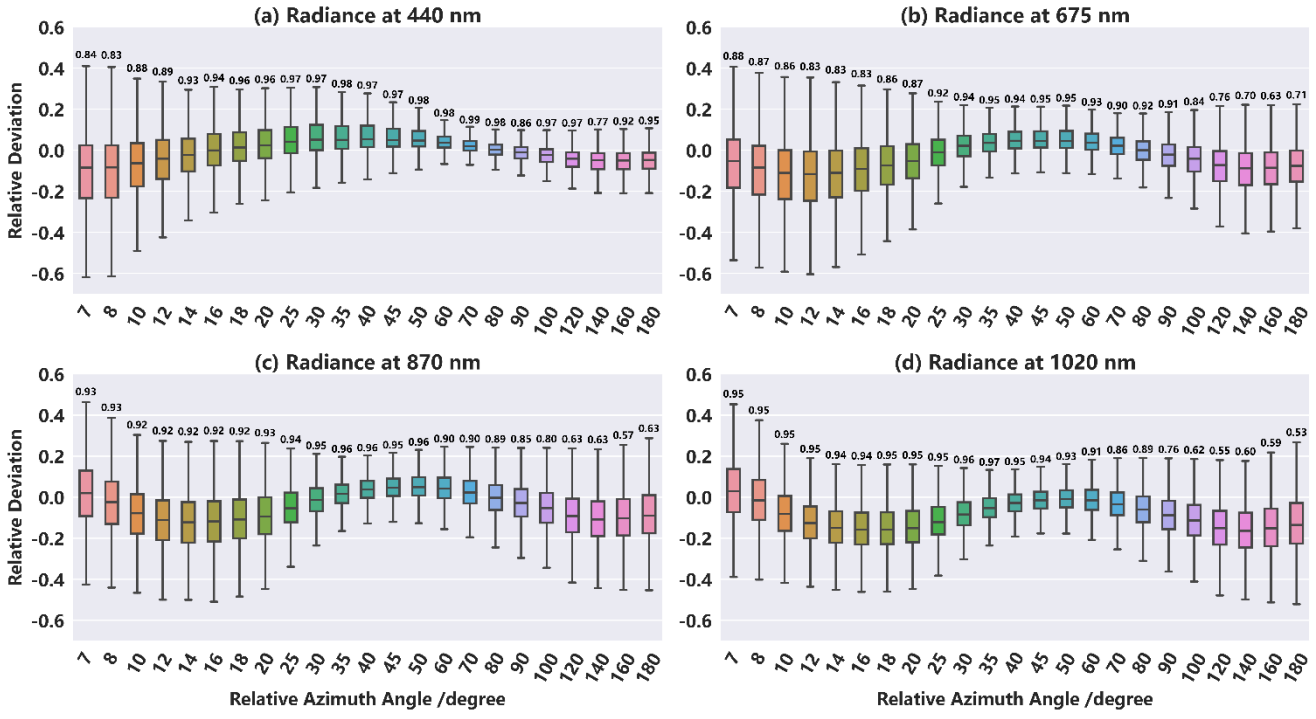


Figure 7. Relative deviation between radiance simulated from EML-based retrieval results and photometer observations. Box colors indicate different RAAs, and the numbers inside each box show the corresponding correlation coefficient.

4 Summary

This study presents a novel aerosol retrieval algorithm based on an EML model to infer both optical and microphysical properties from ground-based Sun–sky photometer measurements. The algorithm simultaneously retrieves four key parameters—SSA and g at four observation wavelengths, as well as r_{eff} and FMF—achieving accuracy comparable to that of the AERONET official algorithm and products. Compared with traditional numerical inversion methods, the EML-based algorithm offers three major advantages: it is five orders of magnitude faster by avoiding iterative radiative transfer calculations; it does not rely on prior assumptions or smoothing constraints; and it eliminates convergence issues inherent in statistical optimization methods, reducing missing data caused by non-convergence.

Our EML model is trained on data generated from forward radiative transfer simulations using a combination of T-matrix and VLIDORT models, independent of existing inversion algorithm products and instrument measurements with errors. The

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

simulations span a comprehensive range of aerosol types and atmospheric conditions, ensuring the model's universality and portability. Systematic and propagation errors were evaluated, yielding total retrieval uncertainties of 0.03 for SSA, 0.02 for g, 0.08 for r_{eff}, and 0.09 for FMF. Application to raw photometer measurements demonstrates strong agreement with AERONET products in both retrieved parameters and optical residuals. SHAP-based feature importance analysis verifies the physical interpretability of the model: SSA retrieval shows a stronger dependence on AOD compared to the other retrieved parameters, while g retrieval is primarily influenced by sky diffuse radiance across all observation wavelengths. Auxiliary observation geometry also plays a critical role. Finally, error analysis indicates that measurements with RAAs in the range 20 °–100 °and higher AOD values provide more favorable conditions for accurate aerosol retrieval.

Despite these promising results, certain limitations remain. The EML model occasionally produces physically unrealistic values, such as SSA exceeding 1 or g falling below 0; currently, these anomalies are handled through value truncation, which is a practical but suboptimal solution. Moreover, the algorithm presently retrieves only r_{eff} and FMF, without providing full aerosol size distributions or complex refractive index information. Nevertheless, our results highlight the substantial potential of machine learning approaches for addressing ill-posed and nonlinear retrieval problems. Looking forward, ongoing advances in artificial intelligence, coupled with increasingly comprehensive ground-based and satellite observations, are expected to facilitate the development of next-generation aerosol retrieval algorithms and products.

Appendix A: Optical Residual of 440nm

According to the method described in Sect. 2.4, we calculated the residuals for each individual wavelength, using the same plotting approach as in Fig. 6. At 440 nm, our inversion algorithm exhibits smaller residuals. Moreover, the differences between the residuals of the two algorithms, as well as the spatial pattern of r_{eff}, are more pronounced at this wavelength.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

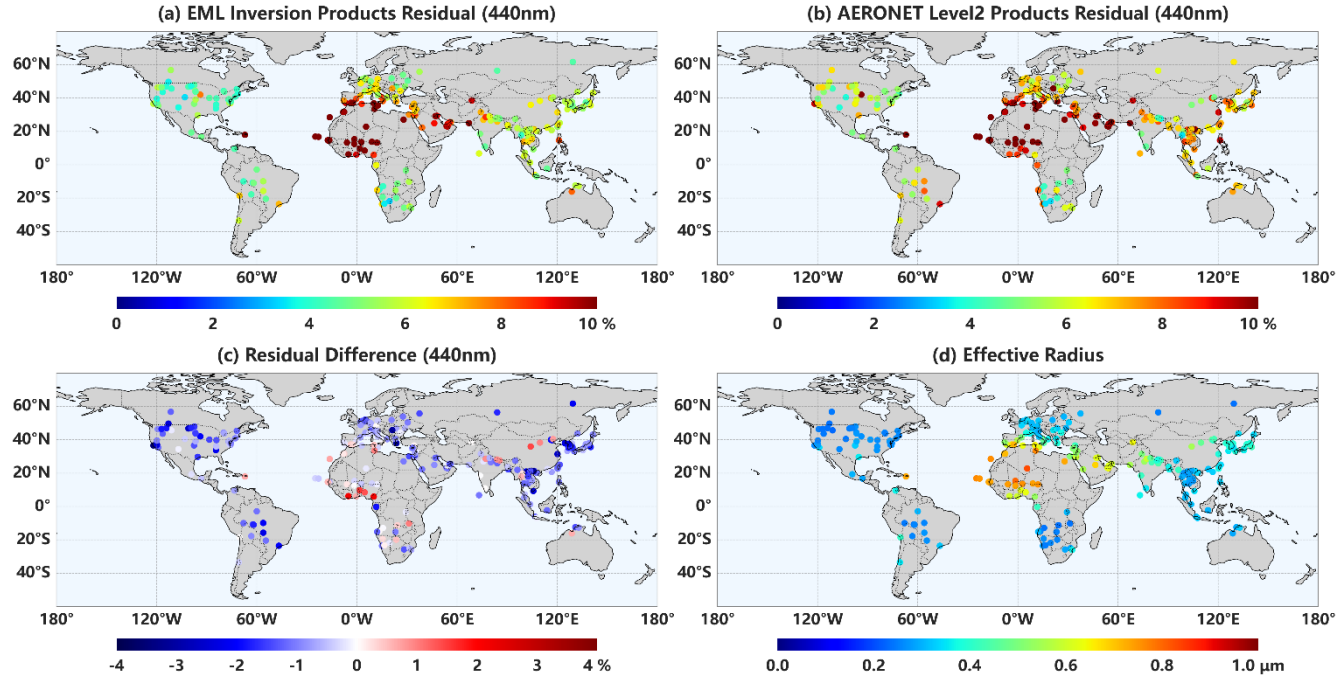


Figure A1. Optical residual at 440 nm of our EML-based and AERONET official inversion algorithms on the testing set. The method is the same as Fig. 6, with only the shortest wavelength (440 nm) selected for radiance.

Appendix B: Application of the EML-Based Retrieval Algorithm to Low-AOD Photometer Observations with Level 1.5 Inversion Products

We applied our EML-based aerosol retrieval algorithm to raw sky photometer observations with low AOD (< 0.4), and the inversion results are shown in Figure B1. This dataset comprises 87,144 cases, none of which have corresponding AERONET level 2.0 inversion products. Compared with the results in Fig. 3, these retrievals exhibit larger deviations from the AERONET level 1.5 inversion products, particularly for SSA and FMF (Fig. B1). However, applying an additional filter to select cases with 440 nm AOD > 0.3 improves the agreement between the two datasets, as illustrated in Fig. B2.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

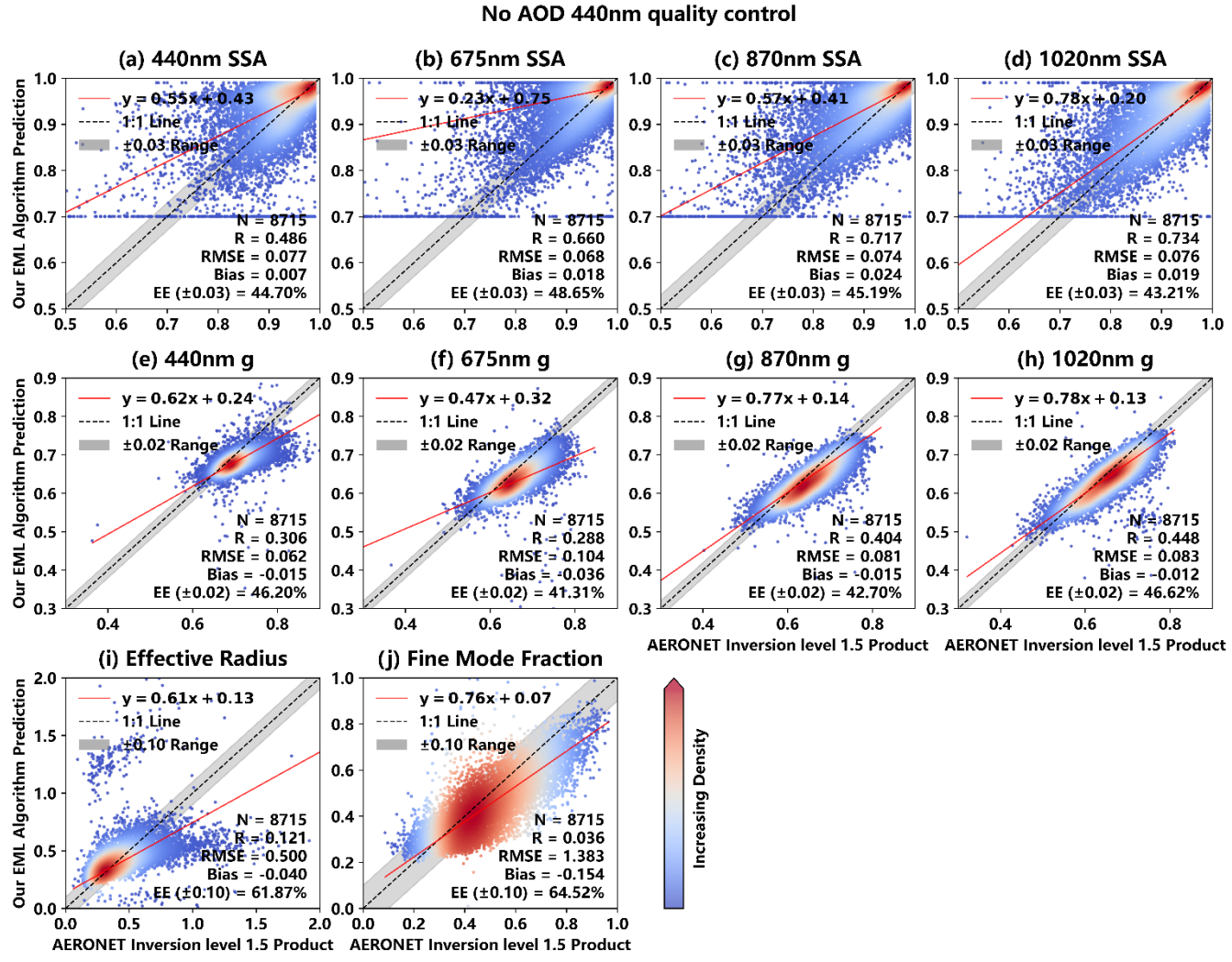


Figure B1. Aerosol parameters retrieved by the EML-based inversion algorithm compared with AERONET Level 1.5 inversion products. All cases correspond to 440 nm AOD < 0.4. The configuration is the same as in Fig. 2. This dataset comprises 81,744 raw Sun-sky photometer measurements, and the scatter points have been thinned to one tenth for clarity.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

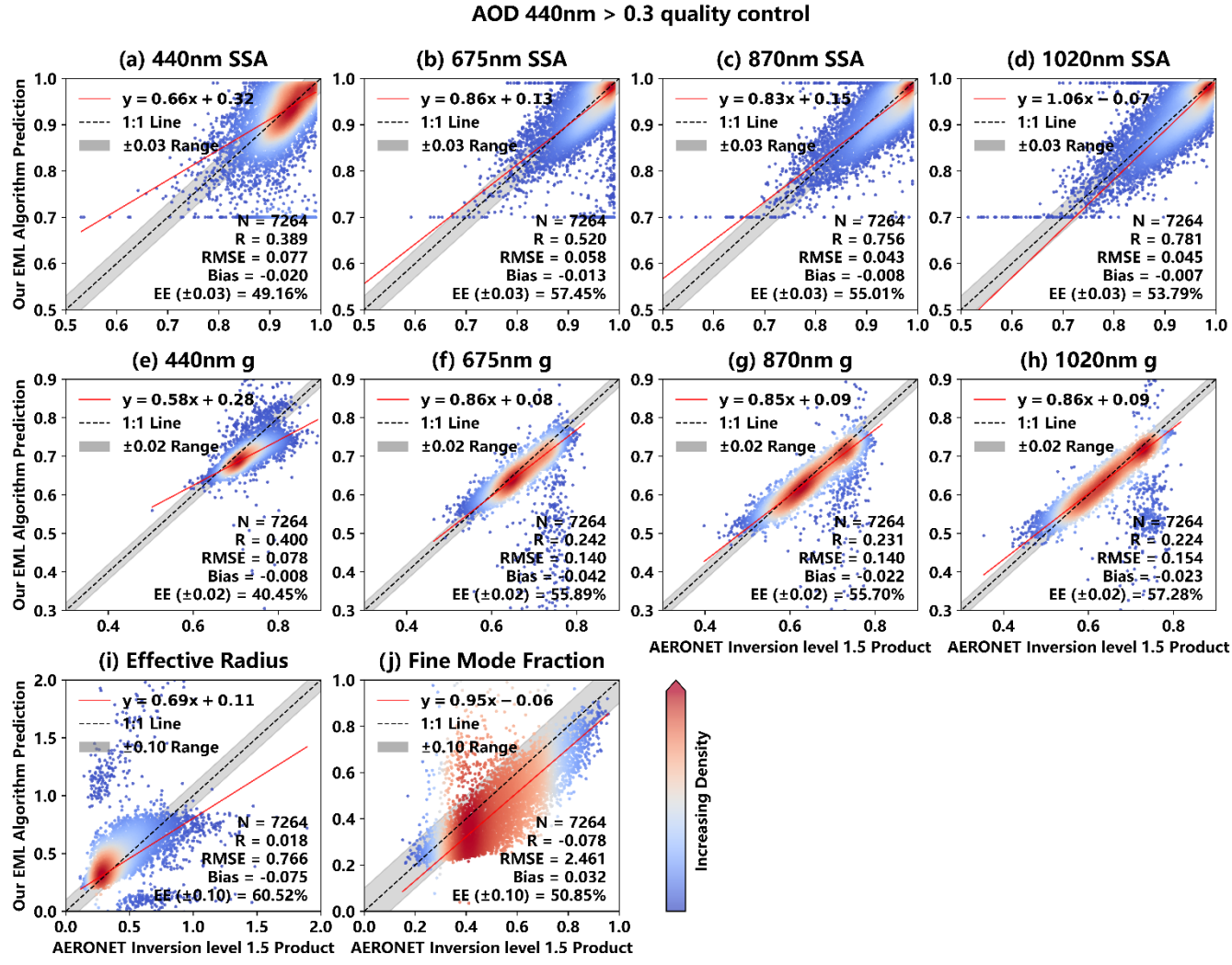


Figure B2. Aerosol parameters retrieved by the EML-based inversion algorithm compared with AERONET Level 1.5 inversion products. All cases correspond to 440 nm AOD between 0.3 and 0.4. The configuration is the same as in Fig. 2. This dataset comprises 7,264 raw Sun–sky photometer measurements, and the scatter points have been thinned to one tenth for clarity.

To further examine retrieval accuracy under varying aerosol loading conditions, we calculated the optical residuals for these 87,144 low-AOD cases and combined them with the 132,067 cases in the testing set (Fig. 2, 440 nm AOD > 0.4). The residuals were grouped according to 440 nm AOD, with the horizontal axis in Fig. B3 binned in intervals of 0.1. The results indicate that when AOD is below 0.4, residuals are significantly higher than for cases with AOD > 0.4. Within the intermediate range of 0.3–1.5, residuals decrease monotonically as AOD increases. At both extremes of the AOD spectrum, retrieval uncertainties

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

tend to rise: low AOD corresponds to weak aerosol signals, which limit retrieval accuracy, whereas high AOD involves more complex aerosol mixtures, increasing inversion uncertainty.

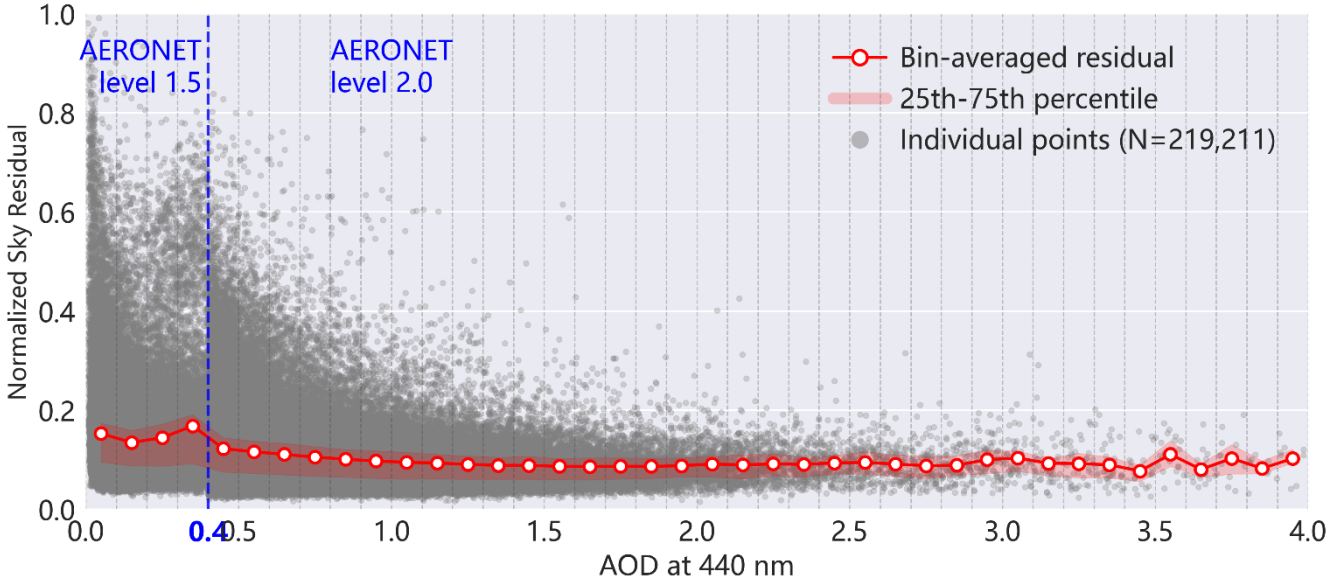


Figure B3. Optical sky residuals binned by 440 nm AOD. Scatter points represent individual cases inverted using the EML-based aerosol retrieval algorithm from raw AERONET site photometer measurements. The vertical dashed line at 440 nm AOD = 0.4 indicates a commonly used quality-control threshold for selecting AERONET Level 2.0 inversion products.

Author contributions. JL and QL conceptualized and designed the study. QL carried out the algorithm development and result analysis, with contributions from JL, ZS, ML, HC, and YZ. QL and JL wrote the initial draft. All authors participated in reviewing and editing the manuscript. JL and YZ oversaw the research and secured funding.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors sincerely thank all personnel involved in the operation and maintenance of AERONET site photometers, as well as the developers and maintainers of the AERONET aerosol inversion algorithms and products. Their efforts have provided invaluable data that made this research possible. The authors thank the editor and anonymous reviewers, who helped improve the manuscript substantially.

Financial support. This study was funded by the National Natural Science Foundation of China (Grant Nos. 42425503 and 42375188).

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

References

Andrews, E., Ogren, J. A., Kinne, S., and Samset, B.: Comparison of AOD, AAOD and column single scattering albedo from AERONET retrievals and in situ profiling measurements, Atmos. Chem. Phys., 17, 6041–6072, https://doi.org/10.5194/acp-17-6041-2017, 2017.

Armstrong, B. H.: Spectrum line profiles: The Voigt function, J. Quant. Spectrosc. Radiat. Transfer, 7, 61–88, https://doi.org/10.1016/0022-4073(67)90057-X, 1967.

Bohren, C. F., and Singham, S. B.: Backscattering by nonspherical particles: a review of methods and suggested new approaches, J. Geophys. Res.-Atmos., 96, 5269–5277, https://doi.org/10.1029/90JD01138, 1991.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cao, M., Zhang, M., Su, X., and Wang, L.: A two-stage machine learning algorithm for retrieving multiple aerosol properties over land: Development and validation, IEEE Trans. Geosci. Remote Sens., 61, 1–17, https://doi.org/10.1109/TGRS.2023.3307934, 2023.

Cazorla, A., Shields, J. E., Karr, M. E., Olmo, F. J., Burden, A., and Alados-Arboledas, L.: Determination of aerosol optical properties by a calibrated sky imager, Atmos. Chem. Phys., 9, 6417–6427, https://doi.org/10.5194/acp-9-6417-2009, 2009.

Chen, X., Zhao, L., Zheng, F., Li, J., Li, L., Ding, H., Zhang, K., Liu, S., Li, D., and de Leeuw, G.: Neural Network AEROsol Retrieval for Geostationary Satellite (NNAeroG) based on temporal, spatial and spectral measurements, Remote Sens., 14, 980, https://doi.org/10.3390/rs14040980, 2022.

Chu, D. A., Kaufman, Y. J., Ichoku, C., Remer, L. A., Tanré, D., and Holben, B. N.: Validation of MODIS aerosol optical depth retrieval over land, Geophys. Res. Lett., 29, MOD2-1–MOD2-4, https://doi.org/10.1029/2001GL013205, 2002.

Davies, C. N.: Size distribution of atmospheric particles, J. Aerosol Sci., 5, 293–300, https://doi.org/10.1016/0021-8502(74)90063-9, 1974.

Deschamps, P. Y., Bréon, F.-M., Leroy, M., Podaire, A., Bricaud, A., Buriez, J.-C., and Sèze, G.: The POLDER mission: Instrument characteristics and scientific objectives, IEEE Trans. Geosci. Remote Sens., 32(3), 598–615, https://doi.org/10.1109/36.297978, 1994.

Dong, Y., Li, J., Zhang, Z., Zheng, Y., Zhang, C., and Li, Z.: Machine learning-based retrieval of aerosol and surface properties over land from the Gaofen-5 Directional Polarimetric Camera measurements, IEEE Trans. Geosci. Remote Sens., 62, 1–15, https://doi.org/10.1109/TGRS.2024.3419169, 2024.

Dubovik, O., and King, M. D.: A flexible inversion algorithm for retrieval of aerosol optical properties from sun and sky radiance measurements, J. Geophys. Res.-Atmos., 105, 20673–20696, https://doi.org/10.1029/2000JD900282, 2000

Dubovik, O., Smirnov, A., Holben, B. N., King, M. D., Kaufman, Y. J., Eck, T. F., and Slutsker, I.: Accuracy assessments of aerosol optical properties retrieved from AERONET sun and sky radiance measurements, J. Geophys. Res.-Atmos., 105, 9791–9806, https://doi.org/10.1029/2000JD900040, 2000.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Dubovik, O., Holben, B. N., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanré, D., and Slutsker, I.: Variability of absorption and optical properties of key aerosol types observed in worldwide locations, J. Atmos. Sci., 59, 590–608, https://doi.org/10.1175/1520-0469, 2002.

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.

Dubovik, O., Herman, M., Holdak, A., Lapyonok, T., Tanré, D., Deuzé, J.-L., et al.: Statistically optimized inversion algorithm for enhanced retrieval of aerosol properties from spectral multi-angle polarimetric satellite observations, Atmos. Meas. Tech., 4, 975–1018, https://doi.org/10.5194/amt-4-975-2011, 2011.

Dutton, E. G., Reddy, P., Ryan, S., and DeLuisi, J. J.: Features and effects of aerosol optical depth observed at Mauna Loa, Hawaii: 1982–1992, J. Geophys. Res.-Atmos., 99(D4), 8295–8306, https://doi.org/10.1029/93JD03520, 1994.

Eck, T. F., Holben, B. N., Reid, J. S., Dubovik, O., Kinne, S., Smirnov, A., O'Neill, N. T., and Slutsker, I.: The wavelength dependence of the optical depth of biomass burning, urban and desert dust aerosols, J. Geophys. Res.-Atmos., 104, 31333–31350, https://doi.org/10.1029/1999JD900923, 1999.

El-Nadry, M., Li, W., El-Askary, H., Awad, M. A., and Mostafa, A. R.: Urban health related air quality indicators over the Middle East and North Africa countries using multiple satellites and AERONET data, Remote Sens., 11, 2096, https://doi.org/10.3390/rs11182096, 2019.

Fan, R., Ma, Y., Jin, S., Gong, W., Liu, B., Wang, W., Li, H., and Zhang, Y.: Validation, analysis, and comparison of MISR V23 aerosol optical depth products with MODIS and AERONET observations, Sci. Total Environ., 856, 159117, https://doi.org/10.1016/j.scitotenv.2022.159117, 2023.

García, O. E., Díaz, J. P., Expósito, F. J., Díaz, A. M., Dubovik, O., Derimian, Y., Dubuisson, P., and Roger, J. C.: Shortwave radiative forcing and efficiency of key aerosol types using AERONET data, Atmos. Chem. Phys., 12, 5129–5145, https://doi.org/10.5194/acp-12-5129-2012, 2012.

Giles, D. M., Sinyuk, A., Sorokin, M. G., Schafer, J. S., Smirnov, A., Slutsker, I., Eck, T. F., Holben, B. N., Lewis, J. R., Campbell, J. R., Welton, E. J., Korkin, S. V., and Lyapustin, A. I.: Advancements in the Aerosol Robotic Network (AERONET) Version 3 database – automated near-real-time quality control algorithm with improved cloud screening for sun photometer aerosol optical depth (AOD) measurements, Atmos. Meas. Tech., 12, 169–209, https://doi.org/10.5194/amt-12-169-2019, 2019.

Hansen, J. E., and Travis, L. D.: Light scattering in planetary atmospheres, Space Sci. Rev., 16, 527–610, https://doi.org/10.1007/BF0016806, 1974.

Hasekamp, O. P., and Landgraf, J.: Linearization of vector radiative transfer with respect to aerosol properties and its use in satellite remote sensing, J. Geophys. Res.-Atmos., 110(D4), D04S12, https://doi.org/10.1029/2004JD005260, 2005.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET – A federated instrument network and data archive for aerosol characterization, Remote Sens. Environ., 66, 1–16, https://doi.org/10.1016/S0034-4257(98)00031-5, 1998.

Hornik, K., Stinchcombe, M., and White, H.: Multilayer feedforward networks are universal approximators, Neural Netw., 2, 359–366, https://doi.org/10.1016/0893-6080(89)90020-8, 1989.

Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., and Feng, Y.: Revealing drivers of haze pollution by explainable machine learning, Environ. Sci. Technol. Lett., 9, 112–119, https://doi.org/10.1021/acs.estlett.1c00865, 2022.

Huttunen, J., Kokkola, H., Mielonen, T., Mononen, M. E. J., Lipponen, A., Reunanen, J., Lindfors, A. V., Mikkonen, S., Lehtinen, K. E. J., Kouremeti, N., Bais, A., Niska, H., and Arola, A.: Retrieval of aerosol optical depth from surface solar radiation measurements using machine learning algorithms, non-linear regression, and a radiative transfer-based look-up table, Atmos. Chem. Phys., 16, 8181–8191, https://doi.org/10.5194/acp-16-8181-2016, 2016.

Hsu, N. C., Lee, J., Sayer, A. M., Kim, W., Bettenhausen, C., and Tsay, S. C.: VIIRS deep blue aerosol products over land: Extending the EOS long-term aerosol data records, J. Geophys. Res.-Atmos., 124, 4026–4053, https://doi.org/10.1029/2018JD029688, 2019.

Kahn, R. A., Gaitley, B. J., Martonchik, J. V., Diner, D. J., Crean, K. A., and Holben, B.: Multiangle Imaging Spectroradiometer (MISR) global aerosol optical depth validation based on 2 years of coincident Aerosol Robotic Network (AERONET) observations, J. Geophys. Res.-Atmos., 110, D10, https://doi.org/10.1029/2004JD004706, 2005.

Kalashnikova, O. V., and Sokolik, I. N.: Modeling the radiative properties of nonspherical mineral dust aerosols, Atmos. Meas. Tech., 6(8), 2131–2154, https://doi.org/10.5194/amt-6-2131-2013, 2013.

Kokhanovsky, A. A. (Ed.): Light Scattering Reviews 8: Radiative Transfer and Optical Properties of Atmosphere and Underlying Surface, Springer-Verlag, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-32106-1, 2013.

Levy, R. C., Remer, L. A., Kleidman, R. G., Mattoo, S., Ichoku, C., Kahn, R., and Eck, T. F.: Global evaluation of the Collection 5 MODIS dark-target aerosol products over land, Atmos. Chem. Phys., 10, 10399–10420, https://doi.org/10.5194/acp-10-10399-2010, 2010.

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmos. Meas. Tech., 6(11), 2989–3034, https://doi.org/10.5194/amt-6-2989-2013, 2013.

Liang, T., Sun, L., and Li, H.: MODIS aerosol optical depth retrieval based on random forest approach, Remote Sens. Lett., 12, 179–189, https://doi.org/10.1080/2150704X.2020.1842540, 2020.

Logothetis, S.-A., Salamalikis, V., and Kazantzidis, A.: The impact of different aerosol properties and types on direct aerosol radiative forcing and efficiency using AERONET version 3, Atmos. Res., 250, 105343, https://doi.org/10.1016/j.atmosres.2020.105343, 2021.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., and Niu, X.: Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning, Electron. Commer. Res. Appl., 31, 24–39, https://doi.org/10.1016/j.elerap.2018.08.002, 2018.

Mao, Q., Zhang, H., Chen, Q., Huang, C., and Yuan, Y.: Satellite-based assessment of direct aerosol radiative forcing using a look-up table established through AERONET observations, Infrared Phys. Technol., 102, 103017, https://doi.org/10.1016/j.infrared.2019.103017, 2019.

Mishchenko, M. I., Liu, L., Travis, L. D., and Lacis, A. A.: Scattering and radiative properties of semi-external versus external mixtures of different aerosol types, J. Quant. Spectrosc. Radiat. Transfer, 88, 139–147, https://doi.org/10.1016/j.jqsrt.2003.12.032, 2004.

Mishchenko, M. I., and Travis, L. D.: T-matrix computations of light scattering by large spheroidal particles, Opt. Commun., 109, 16–21, https://doi.org/10.1016/0030-4018(94)90731-5, 1994.

Mishchenko, M. I., Travis, L. D., Kahn, R. A., and West, R. A.: Modeling phase functions for dustlike tropospheric aerosols using a shape mixture of randomly oriented polydisperse spheroids, J. Geophys. Res.-Atmos., 102, 16831–16847, https://doi.org/10.1029/96JD02110, 1997.

Moosmüller, H., Chakrabarty, R. K., and Arnott, W. P.: Aerosol light absorption and its measurement: A review, J. Quant. Spectrosc. Radiat. Transfer, 110(11), 844–878, https://doi.org/10.1016/j.jqsrt.2009.02.035, 2009.

Moosmüller, H., and Sorensen, C. M.: Small and large particle limits of single scattering albedo for homogeneous, spherical particles, J. Quant. Spectrosc. Radiat. Transfer, 204, 250–255, https://doi.org/10.1016/j.jqsrt.2017.09.029, 2018.

Mugnai, A., and Wiscombe, W. J.: Scattering from nonspherical Chebyshev particles I: cross sections, single-scattering albedo, asymmetry factor, and backscattered fraction, Appl. Opt., 25, 1235–1245, https://doi.org/10.1364/ao.25.001235, 1986.

Nakajima, T., Tonna, G., Rao, R., Kaufman, Y., and Holben, B.: Use of sky brightness measurements from ground for remote sensing of particulate polydispersions, Appl. Opt., 35(15), 2672–2686, https://doi.org/10.1364/AO.35.002672, 1996.

Nakajima, T., Campanelli, M., Che, H., Estellés, V., Irie, H., Kim, S.-W., Kim, J., Liu, D., Nishizawa, T., Pandithurai, G., Soni, V. K., Thana, B., Tugjsurn, N.-U., Aoki, K., Go, S., Hashimoto, M., Higurashi, A., Kazadzis, S., Khatri, P., Kouremeti, N., Kudo, R., Marenco, F., Momoi, M., Ningombam, S. S., Ryder, C. L., Uchiyama, A., and Yamazaki, A.: An overview of and issues with sky radiometer technology and SKYNET, Atmos. Meas. Tech., 13(8), 4195–4218, https://doi.org/10.5194/amt-13-4195-2020, 2020.

Omar, A. H., Winker, D. M., Tackett, J. L., Giles, D. M., Kar, J., Liu, Z., Vaughan, M. A., Powell, K. A., and Trepte, C. R.: CALIOP and AERONET aerosol optical depth comparisons: One size fits none, J. Geophys. Res.-Atmos., 118, 4748–4766, https://doi.org/10.1002/jgrd.50330, 2013.

Osborne, S. R., Johnson, B. T., Haywood, J. M., Baran, A. J., Harrison, M. A. J., and McConnell, C. L.: Physical and optical properties of mineral dust aerosol during the Dust and Biomass-burning Experiment, J. Geophys. Res.-Atmos., 113, D00C03, https://doi.org/10.1029/2007JD009551, 2008.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Ott, W. R.: A physical explanation of the lognormality of pollutant concentrations, J. Air Waste Manag. Assoc., 40, 1378–1383, https://doi.org/10.1080/10473289.1990.10466789, 1990.

Qi, L., Liu, R., and Liu, Y.: Retrieval of aerosol single-scattering albedo from MODIS data using an artificial neural network, Remote Sens., 14, 6341, https://doi.org/10.3390/rs14246341, 2022.

She, L., Li, Z., de Leeuw, G., Wang, W., Wang, Y., Yang, L., Feng, Z., Yang, C., and Shi, Y.: Time series retrieval of multi-wavelength aerosol optical depth by adapting Transformer (TMAT) using Himawari-8 AHI data, Remote Sens. Environ., 305, 114115, https://doi.org/10.1016/j.rse.2024.114115, 2024.

Sinyuk, A., Holben, B. N., Eck, T. F., Giles, D. M., Slutsker, I., Korkin, S., Schafer, J. S., Smirnov, A., Sorokin, M., and Lyapustin, A.: The AERONET Version 3 aerosol retrieval algorithm, associated uncertainties and comparisons to Version 2, Atmos. Meas. Tech., 13, 3375–3411, https://doi.org/10.5194/amt-13-3375-2020, 2020.

Spurr, R. J. D.: VLIDORT, a linearized pseudo-spherical vector discrete ordinate radiative transfer code for forward modeling and retrieval studies in multilayer multiple scattering media, J. Quant. Spectrosc. Radiat. Transfer, 102, 316–342, https://doi.org/10.1016/j.jqsrt.2006.05.005, 2006.

Sun, J., Veefkind, J. P., van Velthoven, P., and Levelt, P. F.: Evaluating modelled aerosol absorption by simulating the UV aerosol index using machine learning, in: EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-8878, https://doi.org/10.5194/egusphere-egu2020-8878, 2020.

Tao, M., Chen, J., Xu, X., Man, W., Xu, L., Wang, L., Wang, Y., Wang, J., Fan, M., Shahzad, M. I., and Chen, L.: A robust and flexible satellite aerosol retrieval algorithm for multi-angle polarimetric measurements with a physics-informed deep learning method, Remote Sens. Environ., 297, 113763, https://doi.org/10.1016/j.rse.2023.113763, 2023.

Taylor, M., Kazadzis, S., Tsekeri, A., Gkikas, A., and Amiridis, V.: Satellite retrieval of aerosol microphysical and optical parameters using neural networks: a new methodology applied to the Sahara Desert dust peak, Atmos. Meas. Tech., 7, 3151–3175, https://doi.org/10.5194/amt-7-3151-2014, 2014.

Turner, D. D., Ferrare, R. A., and Brasseur, L. A.: Average aerosol extinction and water vapor profiles over the Southern Great Plains, Geophys. Res. Lett., 28, 4441–4444, https://doi.org/10.1029/2001GL013691, 2001.

van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., and Villeneuve, P. J.: Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application, Environ. Health Perspect., 118, 847–855, https://doi.org/10.1289/ehp.0901623, 2010.

Vucetic, S., Han, B., Mi, W., Li, Z., and Obradovic, Z.: A data-mining approach for the validation of aerosol retrievals, IEEE Geosci. Remote Sens. Lett., 5, 113–117, https://doi.org/10.1109/LGRS.2007.912725, 2008.

Wang, L., Zhao, Y., Shi, J., Ma, J., Liu, X., Han, D., Gao, H., and Huang, T.: Predicting ozone formation in petrochemical industrialized Lanzhou city by interpretable ensemble machine learning, Environ. Pollut., 318, 120798, https://doi.org/10.1016/j.envpol.2022.120798, 2023.

Whitby, K. T.: The physical characteristics of sulfur aerosols, Atmos. Environ., 12, 135–159, https://doi.org/10.1016/0004-6981(78)90196-8, 1978.

https://doi.org/10.5194/egusphere-2025-4936
Preprint. Discussion started: 15 October 2025

Zhao, Y., Wang, L., Luo, J., Huang, T., Tao, S., Liu, J., Yu, Y., Huang, Y., Liu, X., and Ma, J.: Deep learning prediction of polycyclic aromatic hydrocarbons in the High Arctic, Environ. Sci. Technol., 53, 13238–13245, https://doi.org/10.1021/acs.est.9b05000, 2019.

Zhang, L., Wang, L., Ji, D., Xia, Z., Nan, P., Zhang, J., Li, K., Qi, B., Du, R., Sun, Y., Wang, Y., and Hu, B.: Explainable ensemble machine learning revealing the effect of meteorology and sources on ozone formation in megacity Hangzhou, China, Sci. Total Environ., 927, 171295, https://doi.org/10.1016/j.scitotenv.2024.171295, 2024.