© September 2026 | IJIRT | Volume 13 Issue 4 | ISSN: 2349-6002

Evaluation of Threshold-Based and Hybrid Aerosol Classification Frameworks Using AERONET Observations

Varsha Hole[1], Rahul Raut[2], Mohit Jangale[3], Karan Khonde[4], and Pravin Hole[5]

[1,2,3,4] Department of Computer Science and Engineering, Sardar Patel Institute of Technology, Mumbai, Maharashtra 400058, India

[5] Department of Computer Engineering, D. J. Sanghvi College of Engineering, Mumbai, Maharashtra 400056, India

***Abstract*—Accurate classification of aerosol types is crucial for understanding their optical properties and their effect on the Earth–atmosphere system. Aerosol classification is typically performed either by fixed thresholding of optical parameters or, more recently, by data-driven probabilistic approaches such as Gaussian kernel density estimation (GKDE). Previous studies have applied the two approaches independently, and it remains poorly understood how the choice of method affects aerosol-type attribution, representative aerosol characteristics, and subsequent optical interpretation. In this study, two aerosol classification methods are investigated using long-term Aerosol Robotic Network (AERONET) inversion products at 47 global sites. A complete hybrid aerosol classification framework is implemented, comprising Gaussian kernel density-based preliminary classification, selection of representative aerosol observations, generation of a consistent optical database using the Mie scattering model, and label refinement through machine learning with ensemble classifiers. Separately, a traditional threshold-based classification is applied to the same AERONET dataset following fixed-parameter decision criteria; the threshold-based labels do not undergo representative-observation selection, Mie-based optical modelling, or machine learning, and serve only as a reference for comparison. Results from both pathways are analysed to investigate differences in aerosol-type distributions, the redistribution of observations between aerosol categories, variations in representative aerosol observations and their refractive indices, and the resulting effect on the interpretation of aerosol optical characteristics. Despite their fundamentally different design, the two methods produce highly consistent aerosol-type assignments, with agreement exceeding 98 % for all four aerosol types (dust, mixed, urban/industrial, and biomass-burning). The differences that occur are confined to a small fraction of observations near class boundaries, where the fixed-threshold scheme and the density-based hybrid framework treat transitional cases differently. Because label assignments largely coincide, the representative refractive indices and optical properties derived under the two schemes are also closely comparable. The hybrid framework therefore reproduces the established threshold-based classification while adding a physically grounded, probabilistic description of aerosol types, providing a quantitative basis for understanding how method choice affects aerosol-type attribution in climate-relevant analyses.**

***Index Terms*—Aerosol-type classification, Gaussian kernel density estimation, Mie scattering model, optical properties, random forest, threshold-based classification.**

I. INTRODUCTION

The Earth–atmosphere system depends strongly on aerosol interactions. Aerosol particles interact with the Earth's radiation budget and atmosphere via processes involving aerosol–radiation and aerosol–cloud interactions, and through aerosol effects on air quality and on regional to global climate variability [1], [2]. The effect on radiation and climate, in both magnitude and sign, is largely determined by particle type, that is, by particle size, composition, morphology, and absorption properties. Reliable aerosol-type identification is therefore an essential step toward understanding the underlying aerosol sources and physical properties and toward attributing climate effects.

Despite several decades of observational and modelling activity, aerosol classification remains difficult because the population of particles can vary strongly in space and time, and several particle types commonly coexist within a vertical atmospheric column. The composition and optical properties of dust particles originating from desert surfaces, of biomass-burning emissions, and of urban–industrial particles may change significantly during transport, causing the spectral aerosol optical properties of mixed sources to become indistinguishable when based on a limited number of parameters [3], [4], [5]. This phenomenon is especially prominent for transitional or mixed aerosol types, such as those observed in South Asia and over parts of Africa and South America.

Traditionally, classification approaches have been threshold based and require arbitrary thresholds to classify aerosol types using one or a few parameters, such as aerosol optical depth (AOD), Ångström exponent (AE), single scattering albedo (SSA), and fine-mode fraction (FMF) [3], [6], [5], [7]. Such schemes are efficient and can be applied directly to data derived from aerosol retrieval algorithms and from monitoring networks such as AERONET. However, because the threshold values are selected empirically, the classification definitions may become subjective and variable across locations and aerosol regimes, and small changes in an observed parameter can drastically alter aerosol-type attribution, introducing uncertainty.

In recent years, many studies have used data-driven statistical methods such as Gaussian kernel density estimation (GKDE) to classify aerosols. GKDE is a non-parametric probability density estimation technique that can estimate the probability distribution of multidimensional datasets [8]. It assigns observations to specific aerosol types by estimating the density functions of a given dataset in a high-dimensional parameter space; density maxima in that space are used to define cluster centroids and to determine cluster regions. Applications of Gaussian kernel and hybrid clustering approaches have recently been developed using the AERONET dataset and have been successful in identifying specific aerosol types [9].

Nevertheless, threshold-based and Gaussian kernel density-based approaches have been applied independently, and the extent to which the choice of methodology affects aerosol-type designation at the level of the individual observation is not well understood. Whether a strict, parameter-defined boundary or a statistical density profile is considered more appropriate will naturally yield different classifications, and therefore a different selection of representative observations and associated properties. These differences - rather than one type of classification being “better” - have scientific implications for the interpretation of aerosol-type statistics and for the estimation of aerosol climate impact.

The primary goal of the present work is to develop a complete hybrid aerosol classification process, combining Gaussian kernel density estimation as a preliminary classification step, selection of representative observations, generation of a Mie-scattering-based optical database, and refinement using machine learning. Separately, a conventional threshold-based classification is performed on the same AERONET data. The threshold-based scheme does not undergo representative-observation selection, estimation of particle refractive indices, or machine-learning refinement; it serves solely as an analytical tool for direct comparison of labels against the GKDE-based classification.

The study addresses (i) differences in aerosol-type frequency, (ii) how the classification procedure leads to redistribution of observations among aerosol types, (iii) differences in the representative observations of a specific aerosol type and their related optical properties, and (iv) how these differences affect the interpretation of aerosol optical properties. By comparing a purely conventional and a complete methodology, the paper aims to disentangle the impact of the classification procedure from other factors contributing to aerosol-type identification. Importantly, machine learning is used solely for the refinement stage within the complete framework, and no processing beyond threshold-based classification logic is carried out in the comparison methodology, so that methodological effects on aerosol-type allocation are strictly isolated.

II. STUDY REGIONS

Observations from the Aerosol Robotic Network (AERONET), a worldwide system of sun–sky radiometers that acquires and processes quality-assured long-term data on optical and microphysical aerosol properties [10], [11], are employed in the present study. The AERONET network provides reliable and consistent data with global coverage over the full spectrum and is well established for the characterization and classification of aerosols in different climates [12], [3].

For a meaningful comparison of the influence of methodology on aerosol-type identification, this work uses the same 47 AERONET sites as [9]. These sites were originally chosen after an extensive literature search for regions dominated by a distinct aerosol type, so as to minimize uncertainty in their baseline aerosol properties. Using an identical station set eliminates possible uncertainty arising from station selection or from regional aerosol climatic features.

The 47 sites cover six continents - Asia, Africa, Australia, Europe, North America, and South America - and are representative of the major aerosol types, including dust, urban/industrial (U/I), biomass-burning (BB), and mixed type, the last of which is further decomposed into mixed-coarse and mixed-fine [5], [13]. Marine-type sites were not chosen, since their typical aerosol optical depths (< 0.4) would prevent robust inversion of the optical and microphysical properties required by the two classification methods [9].

Dust sites are located mainly over North Africa and West Asia, in close vicinity to large desert source regions such as the Sahara; they exhibit coarser particles, generally weak-to-moderate absorption, and a weak scattering-angle dependence of the Ångström exponent [14], [15]. Urban and industrial aerosol sites are distributed across Europe, North America, and East Asia; they produce fine particles from fossil-fuel combustion, industry, and vehicular emissions, and are characterized by high Ångström exponents and highly variable absorption properties [5], [7].

Biomass-burning aerosol is generally observed over South America and southern Africa, where forest fires and agricultural and savanna burning have a strong influence; such aerosols are rich in organic material and show significant temporal variability in their optical properties according to combustion characteristics and fuel type [3], [16]. In addition, some mixed sites present aerosol characteristics formed by the mixture of different sources following transport and transformation. These sites represent the most challenging case for classification and offer the greatest potential for contrasting the performance of threshold-based and GKDE-based methods.

This set of 47 carefully selected AERONET stations therefore provides ideal test sites for comparing threshold-based and GKDE-based methods with consistent observation characteristics across a wide range of aerosol environments and climate types.

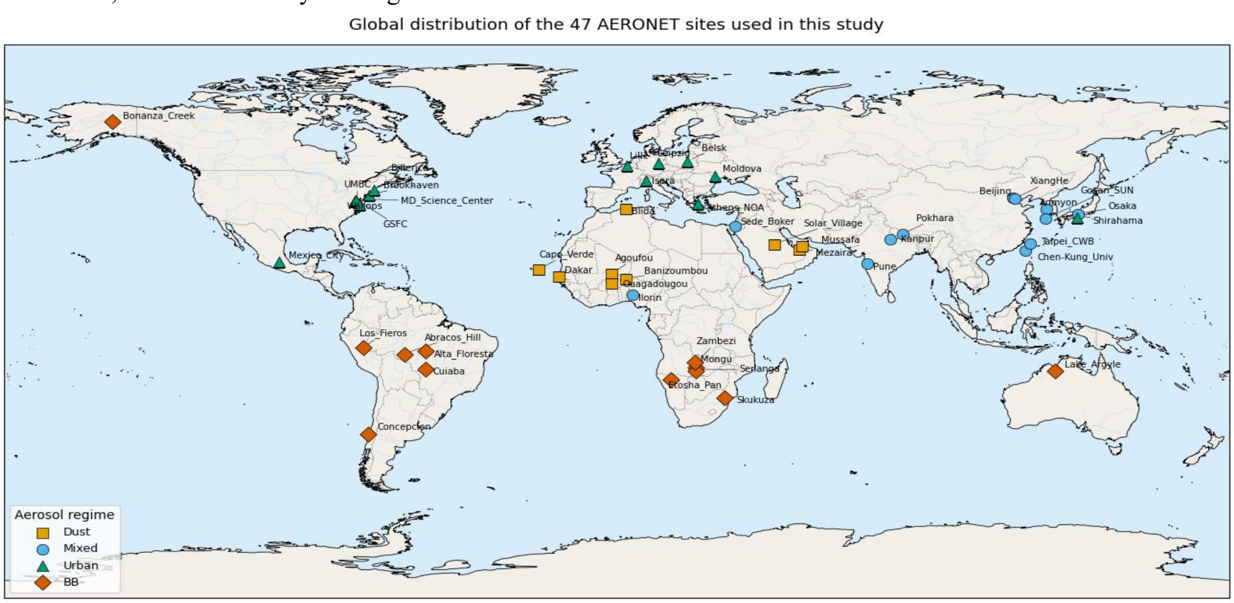

Fig. 1. Study area and the 47 AERONET sites selected by literature review. Site classification follows the dominant aerosol source characteristics reported in the literature and used in the Gaussian kernel density-based framework of [9].

III. DATA AND METHODOLOGY

A. Conceptual Framework and Overall Workflow

Fig. 2. Schematic overview of the proposed aerosol classification and comparison framework.

The characterization and typing of aerosols remains a complex challenge, given that aerosol optical properties change constantly and that aerosol mixing is frequently encountered during transport. To study the effect of the classification method on the determination of aerosol types, two independent classification schemes were applied to the same AERONET dataset, followed by a joint comparative analysis of both sets of results.

The first approach is a fully hybrid aerosol classification scheme modified from the method proposed in [9], which consists of three progressive stages. In Stage 1, first-stage aerosol classification uses Gaussian kernel density estimation (GKDE) to define density maxima of observations, and representative observations near each density maximum are then chosen. In Stage 2, the median values of the complex refractive indices for the estimated aerosol types are determined, and a physically consistent aerosol optical database is generated using a Mie scattering model. In Stage 3, a machine-learning classification is carried out using the optical database generated with the Mie model to derive the final hybrid aerosol types.

The second classification method is a conventional threshold-based scheme that follows the empirical methodology presented in [5], applied directly to the AERONET inversion products with threshold criteria for fine-mode fraction (FMF) and single scattering albedo (SSA). This second method is self-contained and directly yields threshold-based labels of aerosol types, with no selection of representative observations, no optical modelling with Mie theory, and no machine learning.

The two classifications have no methodological or computational connection beyond their reliance on the same AERONET dataset and the same set of 47 AERONET sites. The resulting labels - referred to as “hybrid” and “threshold” - are brought together in the analysis presented in Section III-G, which compares the results in terms of differences in aerosol-type distribution, redistributed data points, and the average characteristics and optical properties of the aerosol types. This study design ensures that any difference in classification results reflects only the difference in classification method. It should be emphasized that machine learning is applied solely in the hybrid scheme; the threshold-based classification is never used to train, validate, or tune any machine-learning algorithm.

B. Data Source and Study Regions

This study uses aerosol information retrieved from AERONET, a ground-based monitoring system of globally distributed sun–sky radiometers that produces long-term, standardized, and quality-assured aerosol optical and microphysical measurements [10]. AERONET sensors acquire measurements of direct solar beam intensity and diffuse sky radiance in the solar spectrum, from which aerosol optical properties can be retrieved using inverse algorithms based on radiative transfer theory [12].

AERONET Version 3 Level 2.0 inversion products were employed to avoid biased retrievals. Level 2.0 data are screened by stringent quality-assurance procedures such as cloud clearing and radiometric calibration and uncertainty estimation, and are therefore applicable to aerosol-type discrimination and statistical analysis [11]. Typically, the uncertainty in retrievals of aerosol optical depth (AOD) is below 0.01 at 440 nm, and the uncertainty in retrievals of SSA is below 0.03 at 500 nm and 550 nm in moderately to highly polluted atmospheres [14]. Observations from the 47 AERONET sites described in Section II, previously analysed by [9], representing dust-dominated, urban/industrial, biomass-burning, and mixed aerosol conditions, are analysed in this study. Marine-dominated sites were excluded because of their usually low AOD, which limits the possibility of reliable inversion retrieval of the aerosol properties needed by both classification schemes.

C. Aerosol Optical Parameters

Aerosol-type classification in this study is based on a set of optical parameters that collectively describe aerosol loading, particle size, and absorption characteristics: aerosol optical depth (AOD), single scattering albedo (SSA), and the extinction Ångström exponent (EAE), all derived from AERONET inversion products. AOD represents the column-integrated extinction of solar radiation by aerosols and is defined as

$$\tau(\lambda) = \int \beta_{ext}(\lambda, z)\, dz \qquad (1)$$

where β_{ext} is the aerosol extinction coefficient and λ denotes wavelength [17]. AOD serves as an indicator of aerosol loading and is commonly used to identify aerosol events such as dust storms and biomass-burning episodes. The extinction Ångström exponent is calculated using AOD values at two wavelengths as

$$EAE = -\,[\ln(\tau_{\lambda 2}) - \ln(\tau_{\lambda 1})] / [\ln(\lambda_2) - \ln(\lambda_1)] \qquad (2)$$

where λ_1 = 440 nm and λ_2 = 870 nm in this study [18]. Lower EAE values are generally associated with coarse-mode aerosols such as mineral dust, whereas higher values indicate fine-mode aerosols from combustion sources [3], [5]. Single scattering albedo is defined as the ratio of scattering to total extinction,

$$SSA = \beta_{sca} / \beta_{ext} \qquad (3)$$

where β_{sca} denotes the scattering coefficient [19]. SSA provides information on aerosol absorption characteristics and plays a critical role in distinguishing between absorbing and non-absorbing aerosol types. These three parameters form the common optical basis from which both the hybrid and threshold pathways are independently derived, ensuring that any differences in the resulting labels reflect methodology rather than differences in the underlying observational inputs.

D. Method A - Hybrid Aerosol Classification Framework

1) Stage 1 - Gaussian kernel density estimation-based preliminary classification:

Gaussian kernel density estimation is a non-parametric statistical technique used to estimate the probability density function of multivariate data without assuming a predefined distribution. In aerosol science, this method has been widely adopted to identify dominant aerosol regimes based on their clustering behaviour in optical parameter space, particularly when aerosol populations overlap due to mixing processes [20], [9].

Selection of optical feature space. Following the methodology of [9], Stage 1 classification was performed in a two-dimensional optical feature space defined by SSA_{440} and $EAE_{440–870}$. SSA_{440} quantifies aerosol absorptivity, distinguishing strongly absorbing aerosols such as biomass-burning and urban/industrial particles from weakly absorbing ones such as dust, while $EAE_{440–870}$ represents particle size characteristics, with low values indicating coarse-mode dominance and high values indicating fine-mode dominance. This SSA–EAE space has been shown to provide strong separability among major aerosol types under diverse atmospheric conditions [3], [21], [9].

Kernel density estimation formulation. Let $\{x_i\}$ for i = 1, …, N represent a set of aerosol observations in the SSA–EAE feature space. The multivariate kernel density estimate at a location x is defined as

$$\hat{f}(x) = (1/N)\, \Sigma\, K_\sigma(x - x_i) \qquad (4)$$

where $\hat{f}(x)$ is the estimated probability density, $K_\sigma(\cdot)$ is the Gaussian kernel function, σ is the bandwidth or smoothing parameter, and N is the total number of samples. The Gaussian kernel function is expressed as

$$K_\sigma(x - x_i) = (1/\sigma)\exp(-\,\lVert x - x_i\rVert^2 / 2\sigma^2) \qquad (5)$$

The choice of a Gaussian kernel ensures smooth density estimation and continuity in regions with sparse sampling, which is particularly important for aerosol datasets influenced by episodic events such as dust storms or biomass-burning outbreaks [22], [9].

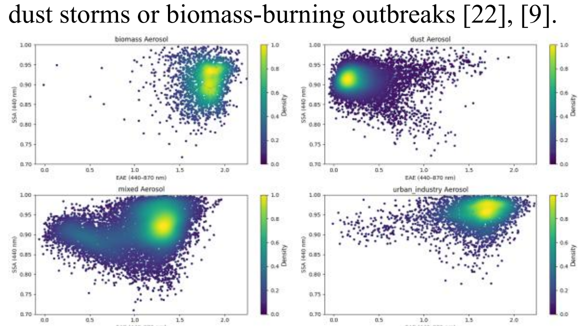

Fig. 3. Clustering distribution of EAE and SSA obtained using the Gaussian kernel density clustering method for the different aerosol types.

Identification of density maxima and aerosol regimes. The estimated density surface exhibits multiple local maxima, corresponding to the most probable SSA–EAE combinations. These maxima represent statistically dominant aerosol populations, interpreted as distinct aerosol regimes. Consistent with the clustering behaviour reported in [9], four major aerosol types emerge from this analysis:

- Dust aerosols, characterized by low EAE values (≈ 0.1–0.4) and moderate SSA (≈ 0.88–0.94), indicating the dominance of coarse-mode mineral particles with weak to moderate absorption.
- Mixed aerosols, occupying an intermediate region in the SSA–EAE space with EAE values typically between ≈ 0.4 and 1.4, representing aerosol populations influenced by both coarse- and fine-mode particles. This category reflects mixtures of mineral dust, anthropogenic pollution, and aged aerosols resulting from long-range transport and atmospheric processing.
- Urban/industrial aerosols, exhibiting high EAE values (> 1.5) and high SSA (> 0.94), corresponding to fine-mode, weakly absorbing aerosols dominated by anthropogenic emissions such as sulfates, nitrates, and industrial pollution.
- Biomass-burning aerosols, characterized by relatively high EAE and a broad range of SSA values, reflecting variability in combustion conditions, fuel type, and aerosol ageing between flaming and smouldering phases.

Unlike threshold-based classification, GKDE does not impose sharp decision boundaries; aerosol regimes are identified probabilistically based on the dominance of local density maxima, allowing natural overlap between neighbouring regimes.

2) Selection of representative aerosol observations:

To build representative aerosol samples for each category identified in Stage 1, a density window around each local maximum was defined, and observations within a fixed radius of that density peak were chosen as high-confidence representative samples of the given aerosol type. Window-based selection restricts interference from transient and mixed aerosol states and ensures that the resulting representative samples are physically meaningful in the context of refractive index determination and optical database generation in Stage 2. It has been confirmed that samples chosen at density maxima show little dispersion in size and refractive index distributions, supporting their value as a representative basis [9]. This representative-observation selection is performed only in the hybrid approach; there is no analogous stage in the threshold-based methodology.

3) Stage 2 - Median refractive index estimation and Mie scattering optical database generation:

Using the representative observations selected above for each aerosol type, a median complex refractive index (CRI) was estimated at each of four AERONET wavelengths (440, 675, 870, and 1020 nm). These median, aerosol-type-specific CRI values, summarized in Table II, serve as a critical physical input to the subsequent Mie scattering simulations, ensuring that the simulated optical properties are anchored to observationally representative absorption and scattering characteristics rather than to arbitrary parameter choices.

Theoretical foundation of Mie scattering**.** Mie scattering theory describes the interaction of electromagnetic radiation with spherical particles of arbitrary size relative to the wavelength of incident radiation [23]. The scattering behaviour is governed by two dimensionless quantities, the size parameter and the complex refractive index, defined respectively as

$$x = 2\pi r / \lambda \tag{6}$$

$$m = n + ik \tag{7}$$

where r is the particle radius, λ is the wavelength, and n and k represent the real and imaginary parts of the refractive index. The real part controls scattering efficiency, while the imaginary part governs absorption. For a given particle size and refractive index, Mie theory yields wavelength-dependent efficiency factors for scattering (Q_{sca}), absorption (Q_{abs}), and extinction (Q_{ext}), related through

$$Q_{ext} = Q_{sca} + Q_{abs} \tag{8}$$

In the present study, wavelength-dependent CRI values were assigned for each aerosol type based on representative values reported in AERONET-based and laboratory studies. The adopted real and imaginary refractive indices at 440, 675, 870, and 1020 nm for dust, mixed, urban/industrial, and biomass-burning aerosols are summarized in Table II.

Aerosol size distribution and extinction coefficient. Atmospheric aerosols are characterized by polydisperse size distributions. Following [9], aerosol populations were represented using lognormal size distributions, expressed as

$$n(r) = [N / (r \ln\sigma_g)] \exp[-(\ln r - \ln r_g)^2 / 2(\ln\sigma_g)^2] \tag{9}$$

where N is the number concentration, r_g is the geometric mean radius, and σ_g is the geometric standard deviation. The extinction coefficient at wavelength λ is obtained by integrating the extinction cross-section over the size distribution,

$$\beta_{ext}(\lambda) = \int \pi r^2 Q_{ext}(r, \lambda, m)\, n(r)\, dr \quad (10)$$

and the corresponding aerosol optical depth is then calculated as

$$AOD(\lambda) = \int \beta_{ext}(\lambda, z)\, dz \quad (11)$$

This formulation ensures that AOD values are physically consistent with aerosol microphysical properties.

Scattering, absorption, and single scattering albedo. Using the Mie-derived efficiency factors, the scattering and absorption coefficients are expressed as

$$\beta_{sca}(\lambda) = \int \pi r^2 Q_{sca}(r, \lambda, m)\, n(r)\, dr \quad (12)$$

$$\beta_{abs}(\lambda) = \int \pi r^2 Q_{abs}(r, \lambda, m)\, n(r)\, dr \quad (13)$$

The single scattering albedo, which quantifies the relative importance of scattering versus absorption, is then given by

$$SSA(\lambda) = \beta_{sca}(\lambda) / \beta_{ext}(\lambda) \quad (14)$$

SSA is a key parameter linking aerosol microphysics to radiative behaviour and is therefore central to aerosol-type discrimination.

Asymmetry parameter**.** The angular distribution of scattered radiation is described by the asymmetry parameter g, defined as the first moment of the scattering phase function P(θ):

$$g(\lambda) = ½ \int P(\theta, \lambda) \cos\theta \sin\theta\, d\theta \quad (15)$$

Values of g close to zero indicate isotropic scattering, whereas larger values correspond to strong forward scattering, typical of coarse-mode aerosols such as mineral dust. The asymmetry parameter is essential for radiative transfer modelling and was therefore included in the generated optical database.

Multi-wavelength optical properties and Ångström exponent. Aerosol optical properties were computed at four wavelengths (440, 675, 870, and 1020 nm) to maintain consistency with AERONET inversion products. Using wavelength-dependent AOD values, the extinction Ångström exponent was derived as

$$EAE_{440–870} = − [\ln AOD(870) − \ln AOD(440)] / [\ln(870) − \ln(440)] \quad (16)$$

This multi-wavelength formulation enables direct comparison between modelled optical properties and observed AERONET retrievals.

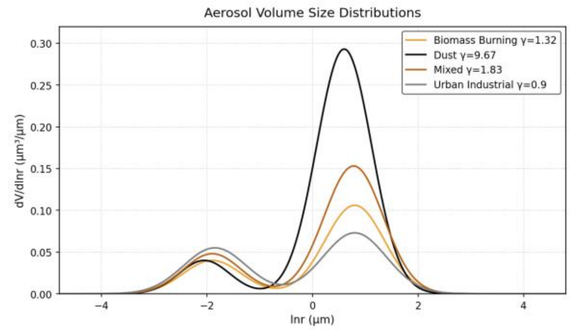

Fig. 4. Volume distribution of the four aerosol types.

Table I. Size distribution parameters of the four aerosol types in coarse and fine modes (unit: µm).

Aerosol type	REFF-fine (µm)	REFF-coarse (µm)	SD-fine (µm)	SD-coarse (µm)
Dust	0.06–0.36	1.19–4.34	0.37–0.94	0.39–0.84
Biomass burning	0.10–0.26	1.32–4.50	0.31–0.74	0.44–0.86
Mixed	0.07–0.44	1.14–4.12	0.31–1.00	0.36–0.88
Urban/industrial	0.10–0.40	1.19–5.43	0.30–0.85	0.41–0.93

Construction of the aerosol optical database**.** The optimal combinations of aerosol size-distribution parameters and complex refractive indices were identified for each aerosol type based on the Stage 1 classification labels. These parameters were used as inputs to the Mie scattering model to generate a physically consistent aerosol optical database containing wavelength-dependent AOD, SSA, asymmetry parameter, and EAE. The resulting optical database served as the input dataset for the supervised machine-learning classification described in Stage 3.

Table II. Real and imaginary parts of the complex refractive index for the four aerosol types at 440/675/870/1020 nm.

Aerosol type	Imaginary index (440/675/870/1020 nm)	Real index (440/675/870/1020 nm)
Dust	0.004346 / 0.001524 /	1.4724 / 1.4893 / 1.4770 / 1.4668

	0.001508 / 0.001612	
Urban/industrial	0.010689 / 0.009231 / 0.009655 / 0.009891	1.4676 / 1.4727 / 1.4754 / 1.4718
Mixed	0.005931 / 0.004084 / 0.004165 / 0.004199	1.4522 / 1.4708 / 1.4751 / 1.4709
Biomass burning	0.017685 / 0.014151 / 0.014325 / 0.014132	1.4964 / 1.5114 / 1.5175 / 1.5136

Dust aerosols have generally low imaginary values, indicating low absorption, whereas biomass-burning aerosols have the highest imaginary values across all wavelengths, reflecting stronger absorption by carbonaceous material. Values for mixed aerosols are intermediate for both parts, consistent with their composite composition. Employing aerosol-type-dependent CRI values generates optical properties that reflect the actual absorption and scattering behaviour of the various aerosol types represented in the AERONET data.

Role of Mie scattering in the present study. The Mie-based optical database is designed to serve three aims: to ensure physical consistency among the aerosol optical parameters inferred from the representative observations in Stage 1; to reduce missing data arising from incomplete AERONET inversion products; and to provide a solid and shared feature space for training the machine-learning model in Stage 3. This stage is applied only in the hybrid pathway; no Mie modelling is performed in the threshold-based classification outlined in Section III-E.

4) Stage 3 - Machine learning-based aerosol classification:

In Stage 3 of the hybrid framework, supervised machine-learning techniques were employed to refine aerosol-type classification using the optical dataset generated through the Mie scattering model in Stage 2. This stage aims to leverage nonlinear relationships among aerosol optical properties to improve classification robustness, particularly in mixed and transitional aerosol regimes where Gaussian kernel density-based classification alone may be insufficient. The machine-learning models were trained using aerosol-type labels originating from the Stage 1 classification, via the representative observations and the resulting Mie-based optical database. Threshold-based labels are not used at any point in this training process. The complete workflow of this stage is illustrated schematically in Fig. 2 and consists of feature engineering, data partitioning, model training and tuning, ensemble learning, and final label generation.

Input dataset and feature construction. The data used to train the models are synthesized from the Mie scattering model and include the wavelength-dependent aerosol optical properties SSA, asymmetry parameter g, AOD, and EAE. Together, these properties comprise a physically consistent and internally related optical feature space for supervised classification. To improve class separation, a number of engineered features were introduced to exploit cross-wavelength dependencies and absorption–particle-size interrelations. These include normalized AODs at longer wavelengths, SSA ratios between short and long wavelengths, AOD ratios, and interaction terms combining EAE and SSA. Such engineered features have been shown to improve the discrimination of aerosol particles, since they represent both spectral gradients and the combined information of size and absorption that individual parameters express only partly. Infinite or undefined values arising in ratio-based features were removed, and missing values were filled using the median of each feature to retain statistical consistency. Feature standardization by Z-scoring was performed to guarantee numerical consistency, which is relevant for scale-sensitive models.

Training–testing strategy and prevention of spatial data leakage. The dataset was split into training and testing partitions at the site level rather than by random sampling, in order to assess generalization ability across geographical locations and to avoid spatial data leakage and the consequent overestimation of accuracy metrics. Observations from a portion of the AERONET sites were used for training and a different set of sites representative of different aerosol regimes was used for testing, mirroring the application of a learned model to a geographic area not represented in the training set. During model training and tuning, GroupKFold cross-validation was used with site names as the grouping variable, preventing observations from the same site from occurring in both training and validation folds.

Handling of class imbalance**.** Aerosol datasets often suffer from imbalance caused by the non-uniform global distribution of aerosol sources and the episodic behaviour of some aerosol phenomena. The Synthetic Minority Over-sampling Technique (SMOTE) was therefore applied during training. SMOTE is a sampling procedure in feature space that creates synthetic observations, producing better class balance without duplicating existing samples. Oversampling was conducted only on the training partition to ensure no leakage into the test set, and it enhanced model sensitivity to minority aerosol classes.

Candidate models and hyperparameter tuning. Three ensemble-based machine-learning algorithms were evaluated for their ability to capture nonlinear relationships and interactions among optical features: Extra Trees (extremely randomized trees), histogram-based gradient boosting, and extreme gradient boosting (XGBoost). Each model was embedded within a unified pipeline consisting of SMOTE-based resampling, feature scaling, and classifier training. Hyperparameters were optimized using randomized search combined with GroupKFold cross-validation to balance computational efficiency and model performance. Model selection was based on cross-validated accuracy on the training set, followed by independent evaluation on the held-out test sites.

Ensemble learning using stacking. To maximize the robustness of classification and to minimize model-dependent errors, a stacking ensemble was implemented using the two best base classifiers. The outputs of the selected base classifiers were fed as features to a random forest meta-model to derive the final aerosol-type prediction. By learning the complementary abilities of the individual classifiers, the stacking model can achieve an increased level of performance and generalization across aerosol conditions. Stacking models were trained with the same resampling, scaling, and site-aware validation as the base models.

Model evaluation and final aerosol-type labelling. To assess model accuracy on a separate spatial sample, all prediction outputs of the stacking ensemble and the base learners on independent AERONET sites were summarized using classification accuracy, confusion matrices, and class-wise metrics (Table III). The results show generally robust generalizability of the stacked model across unseen locations. Following successful validation, the trained stacking pipeline was applied to the full optical dataset to derive the final aerosol-type class labels, which are used as the final classification output for all spatial mapping, statistical calculations, and comparison with the threshold-based technique.

Table III. Performance metrics of the final machine learning-based aerosol classification model across the four aerosol types.

Class	Precision	Recall	F1-score	Support
Biomass burning	0.86	0.88	0.87	792
Dust	0.95	0.95	0.95	2859
Mixed	0.73	0.79	0.76	1520
Urban/industrial	0.88	0.81	0.84	2059
Overall accuracy		0.87		

Role of machine learning in the proposed framework. The machine-learning stage serves as the final integrative step of the hybrid framework. By combining physically grounded optical properties with ensemble learning techniques, this stage enhances aerosol-type discrimination in complex and mixed aerosol environments. Importantly, the classification does not replace physical understanding but builds upon it, refining the aerosol labels generated through the Gaussian kernel density-based classification stage. This hybrid strategy enables reliable aerosol-type mapping across diverse regions and provides a robust foundation for the analytical comparison with threshold-based classification presented in Section III-G.

E. Method B - Threshold-Based Aerosol Classification

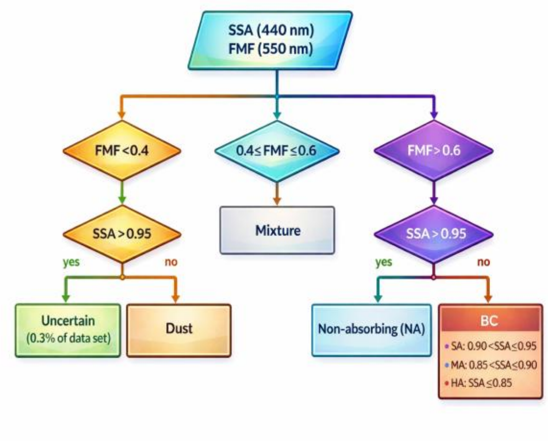

Fig. 5. Flowchart of the threshold-based aerosol classification algorithm for AERONET. HA, MA, SA, and NA represent highly absorbing, moderately absorbing, slightly absorbing, and non-absorbing fine-mode aerosols, respectively.

The threshold-based classification developed in this study is adapted from the empirical approach described by [24]. In that scheme, aerosols are classified on the basis of two directly retrieved AERONET optical parameters - single scattering albedo at 440 nm and fine-mode fraction at 500 nm - using decision-tree logic with predefined thresholds. This branch is applied separately to the identical AERONET data described in Section III-B, and the resulting class labels are employed only for descriptive comparison in Section III-G. No selection of representative observations, Mie calculation, or machine-learning model is applied to the threshold-based classification at any step.

1) Optical parameters used for thresholding:

Mohan et al. [24] evaluated the relative importance of several AERONET-derived optical parameters - SSA, FMF, absorption Ångström exponent, extinction Ångström exponent, and the real and imaginary parts of the refractive index - and found that SSA at 440 nm and FMF at 500 nm carried the highest discriminating power for aerosol-type separation, while the remaining parameters contributed comparatively little additional information. Following this finding, the present threshold-based classification relies exclusively on SSA(440 nm) and FMF(500 nm) as the two governing optical parameters.

2) Determination of dominant particle size using fine-mode fraction:

The fine-mode fraction represents the contribution of fine-mode aerosols to the total aerosol optical depth and is defined as

$$FMF = \tau_{fine} / \tau_{total} \tag{17}$$

where τ_{fine} is the fine-mode aerosol optical depth and τ_{total} is the total aerosol optical depth at 550 nm. Following [24], aerosol observations were first separated by particle size dominance using the criteria FMF > 0.6 for fine-mode-dominated aerosols, FMF < 0.4 for coarse-mode-dominated aerosols, and $0.4 \leq FMF \leq 0.6$ for mixed-mode aerosols. The 0.4–0.6 range is treated as a safety margin that accounts for retrieval uncertainty in the FMF estimate and reflects the natural occurrence of aerosol mixing, a rationale originally proposed by [5] and subsequently adopted by [24].

3) Determination of aerosol absorptivity using single scattering albedo:

Aerosol absorptivity was quantified using the single scattering albedo at 440 nm, defined as in (14), where β_{sca} and β_{ext} represent the scattering and extinction coefficients, respectively. Following [24], a primary SSA threshold of 0.95 was used to separate absorbing from non-absorbing aerosol populations within each size-mode branch. Within the absorbing, black-carbon-dominated population, a further three-level subdivision was applied based on SSA magnitude: low absorption for $0.90 < SSA_{440} < 0.95$, medium absorption for $0.85 < SSA_{440} < 0.90$, and high absorption for $SSA_{440} < 0.85$. This graded subdivision captures the varying degree of light absorption associated with differing black-carbon content and combustion characteristics.

4) Aerosol type assignment logic:

Combining the FMF-based size classification and the SSA-based absorption classification, aerosol observations were assigned to classification flags following the decision-tree logic illustrated in Fig. 5, adapted from [24]. For coarse-mode-dominated observations (FMF < 0.4), SSA > 0.95 was flagged as uncertain - a boundary population in which coarse, weakly absorbing particles could not be unambiguously separated from other coarse-aerosol sources such as sea salt - while $SSA \leq 0.95$ was assigned to the dust class. For fine-mode-dominated observations (FMF > 0.6), SSA > 0.95 was assigned to the non-absorbing (FNA) class, representing weakly absorbing fine-mode aerosols such as sulfate, nitrate, and secondary organic species, while SSA ≤ 0.95 was assigned to the black carbon class and further subdivided into low, medium, and high absorption levels. Observations with 0.4 ≤ FMF ≤ 0.6 were directly assigned to the mixture class, irrespective of SSA, reflecting the transitional nature of these aerosol populations. This decision-tree structure yields six primary classification flags: dust, uncertain, mixture, non-absorbing, and black carbon subdivided into low, medium, and high absorption levels.

5) Mapping of threshold-derived classes to four dominant aerosol types:

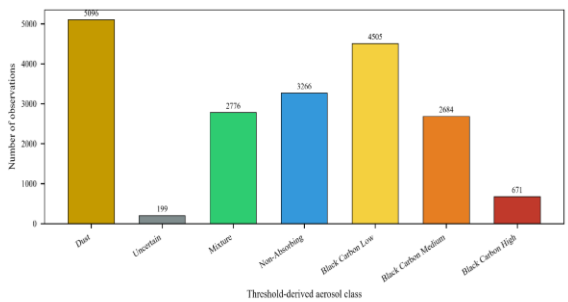

Fig. 6. Distribution of aerosol subtypes obtained from the initial threshold-based classification prior to mapping into the four major aerosol categories.

The threshold-based scheme first partitions the aerosol observations into six classification flags based on the distinction between particle size dominance and absorption strength. For quantitative and conceptual comparison with the four-category typology of the hybrid framework, the classification flags were combined using a physical rationale into four dominant aerosol types: dust, urban/industrial, biomass-burning, and mixed. The two flags associated with strong black-carbon absorption - high and medium - are both mapped onto the biomass-burning classification, since strongly and moderately absorbing carbonaceous aerosols at the surface or within the boundary layer correspond to open burning or biomass sources. The low-absorption black carbon flag is assigned to the urban/industrial category, since it generally corresponds to weakly absorbing aerosol dominated by fossil-fuel combustion or vehicular pollution. The non-absorbing flag is attributed, depending on site type, either to urban/industrial when associated with urban or industrial sites, where it often corresponds to sulfate or nitrate aerosol, or to mixed at non-urban sites, where it may include regionally transported or secondary aerosol formed during transit. Dust remains a separate class, and the mixture flag corresponds to the mixed category. The uncertain category, a borderline case in which coarse aerosol cannot clearly be distinguished from other aerosol classes, represents a very small portion of the dataset and is not considered further. This mapping provides a conceptually consistent correspondence between the threshold-based classes and the four classes of the hybrid framework, enabling meaningful quantitative comparison.

6) Strengths and limitations of the threshold-based approach:

The threshold-based method is attractive because of its computational efficiency, its use of a limited set of directly retrieved optical parameters, and its simple structure, which suits large-scale aerosol climatology studies. It relies only on SSA and FMF, the two most discriminating aerosol parameters identified in [24], excluding unnecessary complexity in higher-dimensional feature spaces and preserving the physical interpretability of decision boundaries. However, like any static-threshold scheme, it is sensitive to the choice of threshold values. Any aerosol measurement near the FMF bounds (0.4 and 0.6) or the SSA bounds (0.85, 0.90, and 0.95) might be reclassified given small variations in aerosol optical properties arising from relative humidity, aerosol ageing, or internal mixing. It can therefore neither adequately address continuous aerosol transitions nor describe the gradual mixing process of aerosols between different species. These considerations motivate the analytical comparison with the Gaussian kernel density-based hybrid framework described in Section III-G.

7) Role of threshold-based classification in the present study:

In this work, the SSA–FMF threshold-based classification scheme of [24] serves strictly as an independent analytical reference. Its final labels are examined alongside the final hybrid labels to investigate methodological sensitivity and data-point redistribution between aerosol categories arising from the two distinct classification philosophies.

F. Summary of Methods A and B

For clarity, the two pathways implemented in this study are summarized as follows. Method A, the hybrid framework, proceeds from AERONET data through Stage 1 (Gaussian kernel density-based classification), selection of representative observations, Stage 2 (median refractive index estimation and Mie scattering optical database generation), and Stage 3 (ensemble machine learning and stacking) to the final hybrid aerosol labels. Method B, the threshold classification, proceeds from AERONET data through threshold classification using FMF and SSA decision rules and a mapping of the resulting flags to four classes, yielding the final threshold-based aerosol labels. The two pathways share only their common input data and optical parameter set; they are otherwise fully independent, and their outputs are brought together only in the analytical framework described next.

G. Analytical Framework for Comparing Hybrid and Threshold-Based Labels

To investigate the influence of classification methodology on aerosol-type attribution, the final hybrid labels and the final threshold-based labels were compared using a dedicated analytical framework comprising five components. First, aerosol-type distribution analysis compares overall aerosol-type frequencies and proportions between the two label sets. Second, label redistribution analysis constructs a confusion and transition matrix quantifying, for each threshold-assigned category, the proportion of observations reassigned to each hybrid-assigned category, and vice versa. Third, representative observation analysis compares the representative observations and associated median refractive indices used within the hybrid framework against the corresponding subsets of threshold-labelled observations, to assess how methodology affects the aerosol optical characteristics attributed to each type. Fourth, spatial and site-wise consistency analysis evaluates aerosol-type composition at individual AERONET sites under both label sets to assess the geographic coherence of each methodology. Fifth, optical property distribution analysis compares AOD, SSA, and EAE distributions associated with each aerosol type under the two label sets.

This analytical framework is applied exclusively at the post-classification stage; it does not feed back into, or alter, either the hybrid or the threshold classification pipeline. Its purpose is to characterize how and where the two methodologies diverge, and to interpret these divergences in terms of the underlying physical and statistical assumptions of each approach, rather than to determine which methodology performs better.

IV. RESULTS AND DISCUSSION

This section presents an analysis of the aerosol classification characteristics obtained from two independently constructed frameworks applied to the same AERONET dataset: a conventional threshold-based classification derived from established decision boundaries, and a hybrid framework comprising Gaussian kernel density estimation, representative observation selection, Mie-scattering-based synthetic optical modelling, and random forest classification. The objective is not to determine which methodology performs better, but to characterize the aerosol populations, physical properties, and spatial patterns that each methodology produces, and to interpret the differences between them in light of established aerosol optical theory [14], [5], [9].

A. Threshold-Based Aerosol Classification Analysis

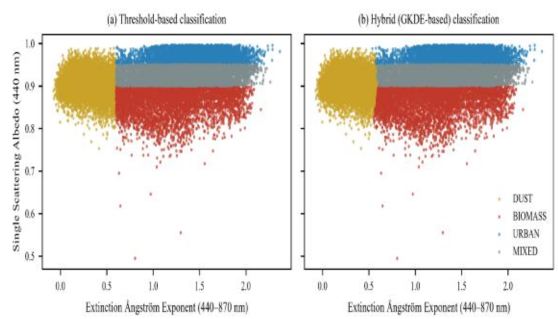

Fig. 7. Threshold and hybrid classification results in SSA–EAE space.

Fig. 7 shows the distribution of aerosol types in SSA–EAE space obtained using the threshold-based approach. The method produces sharp, axis-aligned decision regions, an expected consequence of applying fixed cut-offs on SSA and EAE. This yields a classification that is transparent and computationally efficient, and it performs consistently in regions where aerosol optical properties are well separated - for instance, at low EAE values associated with coarse-mode mineral dust. In intermediate EAE ranges (≈ 0.8–1.5), however, the fixed boundaries partition what is physically a continuous transition zone between aerosol regimes into discrete categories. This is an intrinsic property of any rule-based partitioning of a continuous optical-property space, and it is useful to characterize before comparing it against the density-based structure.

B. Hybrid Framework: Density-Based Representative Aerosol Analysis

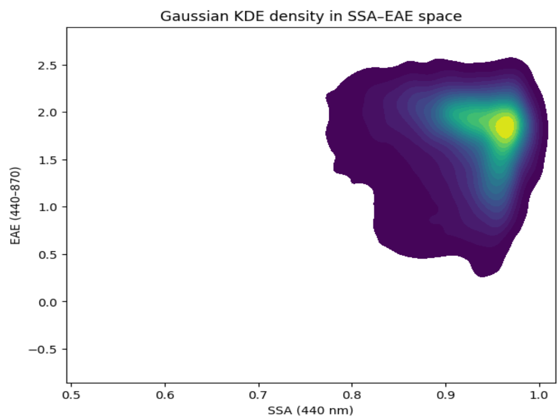

Fig. 8. Gaussian kernel density distribution in SSA–EAE space.

The aerosol distribution resulting from the GKDE stage of the hybrid approach is displayed in Figs. 7 and 8. Instead of relying on predefined regime ranges, GKDE derives the joint probability density of SSA and EAE directly from the data and consequently generates smooth, uninterrupted regime boundaries, designating points at local density maxima as representative samples. Fig. 8 shows distinct density peaks in the SSA–EAE plane, in good agreement with the existence of coarse dust particles, fine anthropogenic particles, and populations of particles in a mixed or aged state.

The dominant density maximum is located at high SSA (≈ 0.90–0.98) and high EAE (≈ 1.6–1.9), corresponding to the most frequent populations of fine, weakly absorbing anthropogenic particles. A broad extension toward low EAE (≈ 0.1–0.4) at moderate SSA (≈ 0.88–0.94) reflects coarse-mode, dust-rich populations, while the intermediate region (EAE ≈ 0.5–1.4) is populated by mixed and aged aerosols with more variable absorption. These density features are broad and overlapping rather than sharply separated. The density surface maintains the overlap between regimes, a process naturally occurring in atmospheric aerosol ensembles undergoing mixing and ageing [14], [9]. These maxima provide the representative samples used to constrain the Mie scattering simulations.

C. Confusion and Transition Analysis: Threshold Versus Hybrid Labels

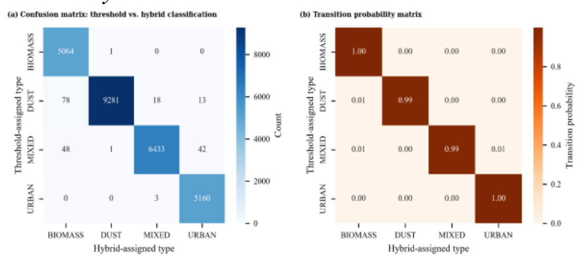

Fig. 9. (a) Confusion matrix and (b) transition probability matrix comparing threshold-based and hybrid (GKDE-based) aerosol classifications.

Fig. 9(a) presents the confusion matrix between threshold-based and hybrid aerosol classifications, quantifying the agreement between the two methods. The matrix is strongly diagonal: the large majority of observations receive the same label under both schemes, with agreement exceeding 98 % for every aerosol type. Agreement is highest for biomass-burning and urban/industrial aerosols, at approximately 100 %, while the small number of off-diagonal cases occurs mainly for dust and mixed observations, where a limited fraction of points is reassigned between neighbouring categories.

The strong consistency across all types reflects the fact that both schemes ultimately partition the same SSA–EAE optical space, so observations with well-separated optical signatures - such as coarse-mode mineral dust with low EAE and moderate SSA - are labelled identically. The few disagreements arise for observations lying close to the threshold boundaries, where small differences in optical properties can move a point between neighbouring categories under the rigid threshold rule but not under the density-based scheme. Fig. 9(b) shows the corresponding transition probabilities between the two label sets.

D. Random Forest Classification on Synthetic and Observed Data

The random forest model trained on the Mie-generated synthetic database exhibits consistent discrimination among aerosol types when applied to real AERONET observations, with feature importance analysis indicating that SSA ratios, AOD ratios, and EAE–SSA interaction terms contribute most strongly to class separation. This indicates that the classes learned from the synthetic optical space remain applicable to observed aerosol optical properties.

E. Aerosol Population Distributions: Threshold Versus Hybrid Framework

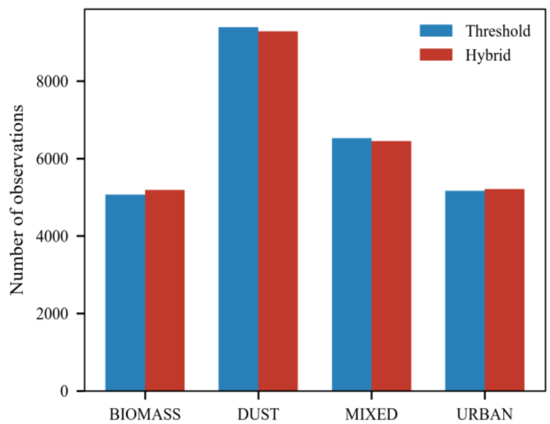

Fig. 10. Comparison of aerosol-type counts derived from the threshold and hybrid methods.

Figs. 9 and 10 show the outputs from the two frameworks in parallel to enable direct comparison. All four aerosol types show a high level of agreement between the frameworks. Agreement is essentially complete for biomass-burning and urban/industrial aerosols and remains above 98 % for dust and mixed aerosols, reflecting the well-separated spectral optical signatures - for example, low EAE and moderate SSA for mineral dust - that both partitioning schemes resolve consistently in SSA–EAE space.

The small discrepancies that remain are concentrated in the transitional observations that overlap in SSA–EAE space, which the two frameworks resolve slightly differently. Almost all threshold-labelled biomass-burning and urban/industrial observations retain the same label under the hybrid method, and about 99 % of mixed observations are also preserved. Dust shows the largest, though still minor, reallocation: about 99 % of threshold-labelled dust observations remain dust, with roughly 1 % redistributed to neighbouring categories. This pattern reflects the differing philosophies of the two schemes - the threshold method assigns a unique label from fixed cut-offs, whereas the density-based hybrid method can move a small number of borderline cases across category boundaries - but in the present dataset the two approaches converge on nearly identical labels.

The relative population counts for each type predicted by the two methodologies are presented in Fig. 10. The counts produced by the two classifications are very similar for every aerosol type: dust is the most abundant class, followed by mixed, urban/industrial, and biomass-burning, and the threshold and hybrid totals differ only marginally within each type. This close correspondence is the expected consequence of the near-identical labelling documented in Fig. 9, and it indicates that the two methodologies handle the transitional regions of the SSA–EAE spectrum in comparable ways for this dataset. Establishing whether one prediction is better than the other would require an independent ground-based study.

F. Site-Wise Aerosol Characteristics and Spatial Patterns

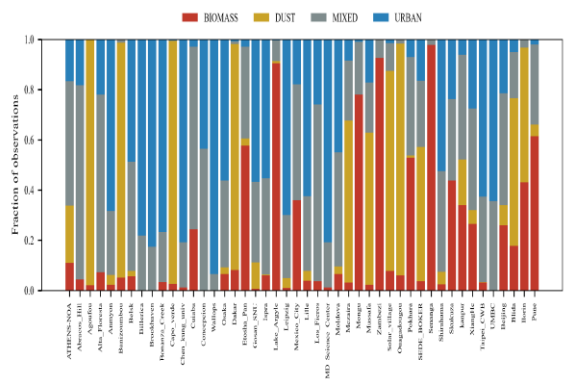

Fig. 11. Site-wise aerosol-type composition based on the final machine learning classification.

Fig. 11 shows the site-wise aerosol-type compositions retrieved from the hybrid, random-forest-refined classification at the 47 AERONET sites used in this work. Dust sites such as Dakar, Banizoumbou, and Capo Verde maintain elevated dust fractions, likely due to their proximity to the main dust sources, while sites primarily affected by anthropogenic and mixed aerosols, such as Kanpur, Xianghe, and Mexico City, remain characterized by significant shares of those types. Sites influenced by biomass burning, such as Alta Floresta, Abracos Hill, and Mongu, tend to show broader aerosol-type compositions, likely due to the variability of burn periods and ongoing aerosol ageing. Overall, this indicates that the hybrid framework robustly captures site-specific and regional differences in aerosol populations.

G. Physical Consistency of Classified Aerosol Types

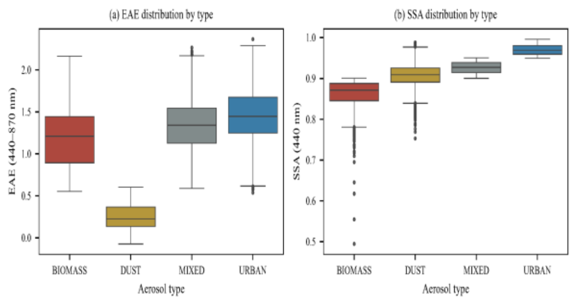

Fig. 12. (a) EAE and (b) SSA distributions by aerosol type.

The distributions of EAE and SSA values for each aerosol class identified by the hybrid classification are plotted in Fig. 12. The dust class generally has low EAE values, reflecting the prevalence of coarse-mode dust aerosols. The urban/industrial class has relatively high SSA values, consistent with weakly absorbing fine-mode aerosols. The biomass-burning class is intermediate, with greater variation in SSA due to variability in burning phase and ageing. The mixed class exhibits the widest range of EAE and SSA values of all classes, consistent with this class encompassing heterogeneous and transient aerosol regimes rather than representing a specific physical aerosol type. These class distributions are in good agreement with those derived from AERONET observational data [14], [5], supporting the physical integrity of the hybrid model-based aerosol types.

Fig. 13 presents the corresponding normalized dominance heatmap across all 47 AERONET sites, showing the fraction of observations assigned to each aerosol type at every site. This reinforces the persistence of the site-specific aerosol characteristics noted above and closely reflects the aggregated SSA–EAE structure identified in Fig. 12.

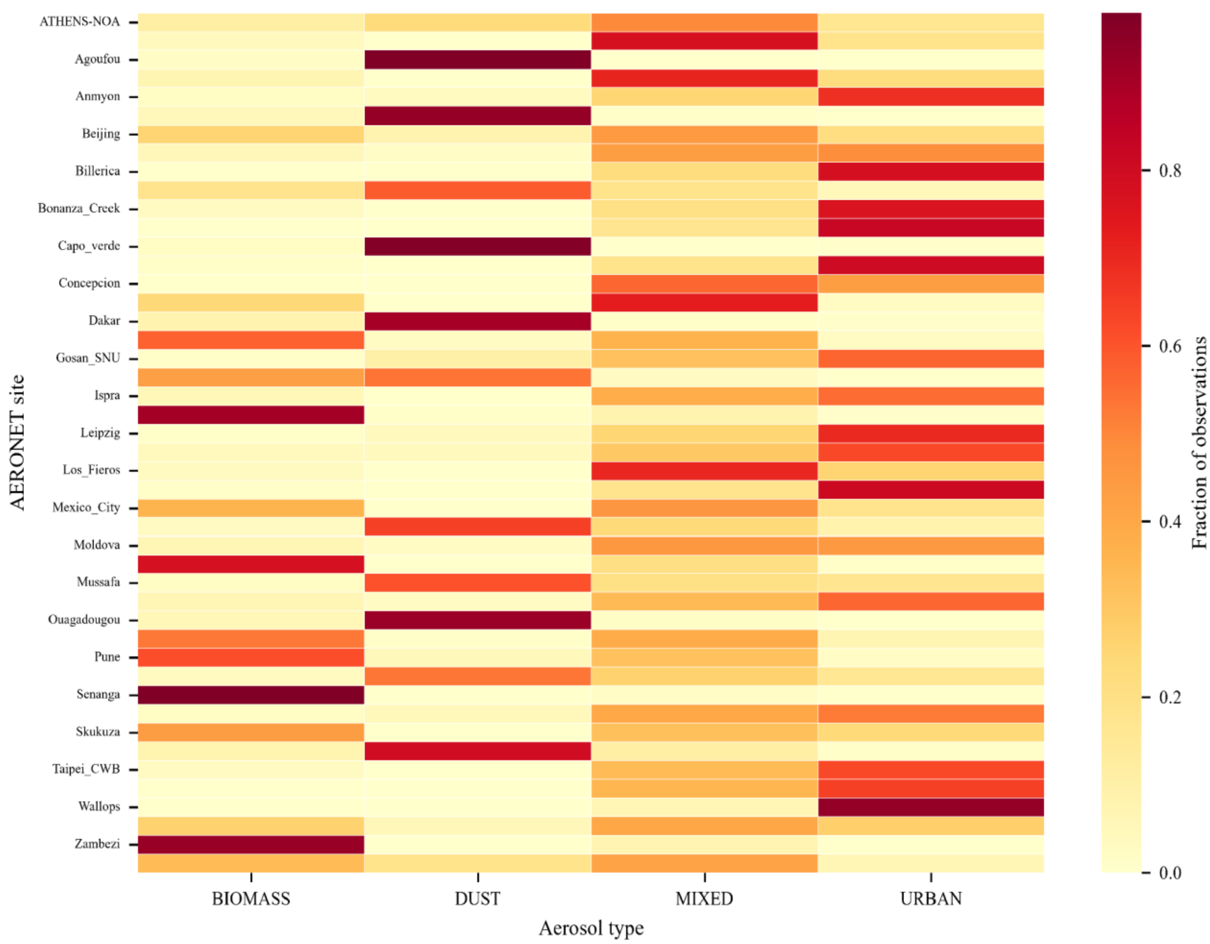

Fig. 13. Site-wise dominance of aerosol types across AERONET stations.

H. Optical Validation Using the Complex Refractive Index

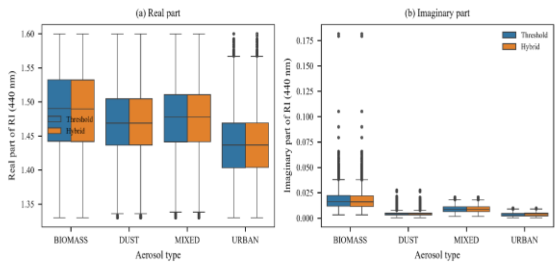

Fig. 14. Refractive index comparison between threshold and hybrid classifications, by aerosol type.

The real and imaginary parts of the complex refractive index under the threshold-based and hybrid classifications are compared per aerosol type in Fig. 14. For every aerosol type, the two parts show closely overlapping distributions under the two classifications, consistent with the high label agreement documented above. The imaginary part, which governs absorption, is smallest for dust and largest for biomass-burning aerosols, and this ordering is reproduced almost identically by both schemes. Because the two methods assign the same label to the large majority of observations, the representative optical properties attributed to each aerosol type are effectively the same, whether class membership is defined by fixed optical limits or by proximity to a local density maximum.

Under a threshold-based scheme, points that lie very close to a given limit, and hence have very similar optical properties, may be assigned entirely to one class or another. In the hybrid framework, such points can instead be associated with the class whose members are local to them in density terms, regardless of absolute limits. This distinction may be relevant when radiation schemes are used to estimate aerosol forcing in the atmosphere.

Fig. 15 presents a spatial extension of the site-specific SSA–EAE classification, showing the clusters of data at example sites that dominate each category of dust, urban/industrial, biomass-burning, and mixed aerosols. These clusters closely reflect the aggregated SSA–EAE structure documented in Figs. 12 and 13, reinforcing the validity of the aerosol types identified.

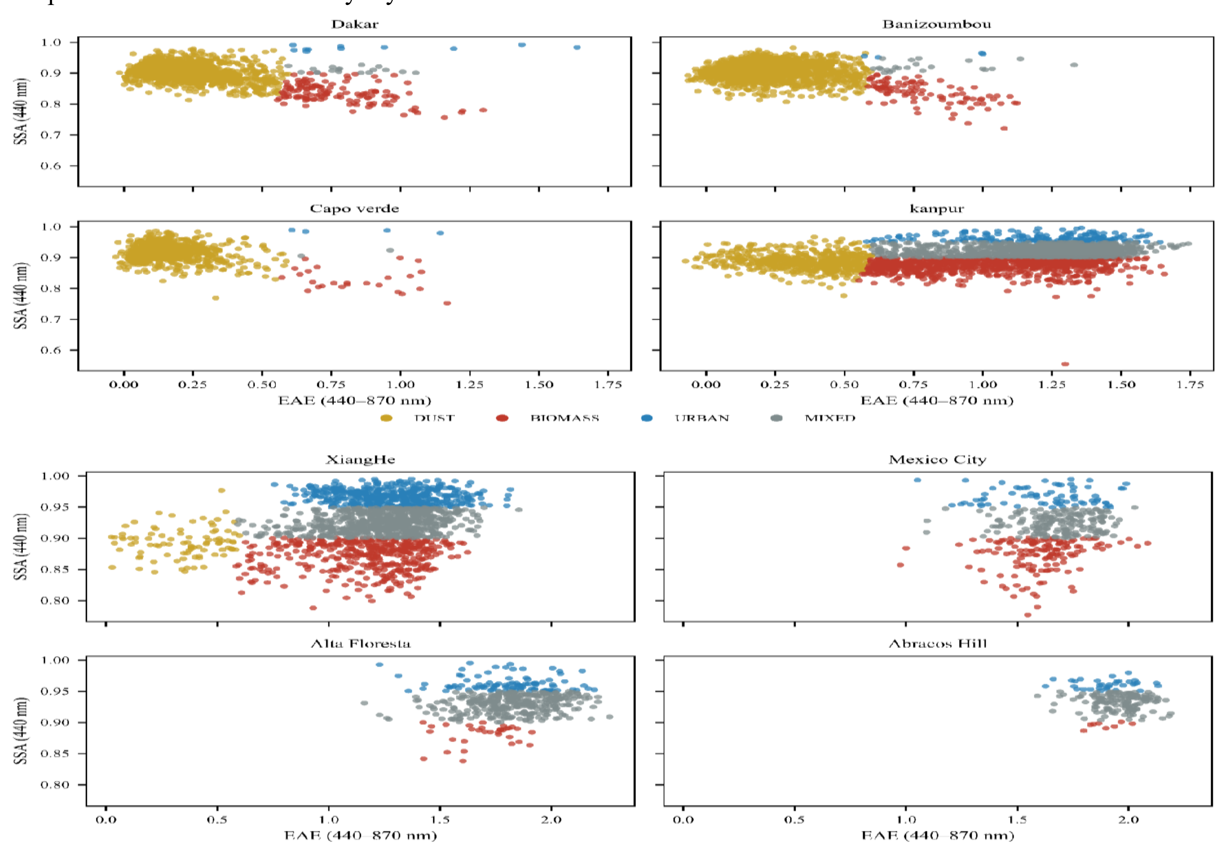

Fig. 15. Composite site-specific SSA–EAE scatter plots for representative sites.

I. Summary of Aerosol Classification Characteristics

Overall, the results define two self-consistent but philosophically divergent characterizations of aerosol types. The threshold-based approach offers a readily interpretable, clearly delineated partition of SSA–EAE space with high within-type accuracy for optically homogeneous aerosol classes, notably mineral dust. However, it forces discontinuities at sharp class transitions within what is a continuum of optical property space. The hybrid framework characterizes the data through the inherently continuous and overlapping structure of the underlying aerosol population distribution, describing class membership probabilistically rather than through fixed cut-offs. In the present dataset this probabilistic description yields labels - and associated complex refractive indices - that closely match those of the threshold scheme, while providing a physically grounded and continuous representation of transitional aerosol regimes. Neither approach can be stated to be inherently more correct than the other; rather, they offer complementary views of the data driven by fundamentally different conceptual bases.

V. CONCLUSIONS

Two classification approaches applied to the same AERONET data were compared: a traditional threshold-based approach and a hybrid approach incorporating Gaussian kernel density estimation, representative observation selection, Mie-scattering-based synthetic optical modelling, and random forest classification. In the first phase of the study, aerosol types were identified independently using the two methodologies: a rule-based thresholding method developed from conventional boundary definitions, and a GKDE-based method that estimates the joint probability density of aerosol optical properties and identifies representative observations at locations of local density maxima. Whereas the threshold-based classification provides clear boundaries across SSA–EAE space, it does not resolve what is physically a continuum of transitional aerosol states, which the GKDE approach maps continuously while identifying representative observations of each density-based aerosol regime.

The two methods were found to produce largely consistent labelling, and the small differences observed are attributed primarily to how each method treats a limited number of transitional observations near the mixed and urban/industrial boundaries. In the second phase, Mie scattering theory was used to build a physically consistent synthetic optical dataset, generating AOD, SSA, asymmetry parameter, and EAE at the four AERONET wavelengths for the four representative aerosol types, based on the selected representative observations and the corresponding complex refractive index and size-distribution pairs from the GKDE-based analysis. The dataset showed good agreement with AERONET inversions and was used to develop and train the classification algorithms. In the third phase, ensemble machine-learning classifiers - Extra Trees, histogram gradient boosting, and XGBoost - were trained on this synthetically generated data using engineered optical features, addressing class imbalance with SMOTE oversampling and site-based information leakage with GroupKFold cross-validation. A stacking classifier with a random forest meta-learner was then used to label real AERONET observations, which formed the inputs for the site-wise and physical consistency analyses.

Several trends are evident from these analyses. Across all aerosol types there is a high level of agreement between the threshold-based and hybrid methods, with more than 98 % of observations receiving the same label under both schemes; the largest, though still minor, disagreements occur for dust and mixed observations near the class boundaries. The complex refractive index distributions obtained under the two schemes are correspondingly similar, so that the hybrid framework reproduces the threshold-based characterization while additionally providing a continuous, physically grounded description of aerosol types. The hybrid framework is considered to produce a better characterization of physical realism, while the clear boundaries provided by the threshold-based approach offer complementary information. Further work would include incorporation of the vertical distribution of aerosols, treatment of the aerosol ageing process, analysis of satellite-based data, and comparison of the accuracy of both methods where sufficient ground-truth information is available for a direct assessment of classification accuracy for each aerosol type.

VI. CODE AND DATA AVAILABILITY

The Python code used for Gaussian kernel density estimation, Mie scattering simulations, and the machine learning classification framework developed in this study is openly available on GitHub at https://github.com/mohitjangale8/AEROSOL_CLASSIFICATION_RESEARCH and has been permanently archived on Zenodo [25]. The AERONET aerosol optical depth and inversion products used in this study are openly available from the NASA AERONET data portal at https://aeronet.gsfc.nasa.gov/.

VII. AUTHOR CONTRIBUTIONS

Varsha Hole: supervision, methodology, project administration, writing - review and editing. Rahul Raut: conceptualization, data curation, formal analysis, investigation, validation, visualization, writing - original draft. Mohit Jangale: conceptualization, methodology, resources, supervision, data curation, formal analysis, investigation, validation, visualization, writing - original draft and review and editing. Karan Khonde: conceptualization, data curation, formal analysis, validation, writing - review and editing. Pravin Hole: supervision, validation, writing - review and editing.

VIII. CONFLICT OF INTEREST

The authors declare that they have no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

IX. ACKNOWLEDGMENT

The authors express their sincere appreciation to the anonymous reviewers for their thoughtful feedback and constructive suggestions. The aerosol observations analysed in this study were sourced from the ground-based Aerosol Robotic Network (AERONET), operated by the National Aeronautics and Space Administration (NASA), and accessed via the AERONET data portal. The dataset includes measurements from 47 NASA-operated AERONET stations located across six continents - Africa, Asia, Australia, Europe, North America, and South America. The authors gratefully acknowledge the efforts of the principal investigators, co-investigators, and technical teams at these stations for their dedication in maintaining and providing high-quality, long-term aerosol measurements for scientific use.

REFERENCES

[1] J. T. Kiehl and B. P. Briegleb, "The relative roles of sulfate aerosols and greenhouse gases in climate forcing," Science, vol. 260, pp. 311–314, 1993.

[2] V. Ramanathan, P. J. Crutzen, J. T. Kiehl, and D. Rosenfeld, "Aerosols, climate, and the hydrological cycle," Science, vol. 294, pp. 2119–2124, 2001.

[3] T. F. Eck, B. N. Holben, J. S. Reid, O. Dubovik, A. Smirnov, N. T. O'Neill, I. Slutsker, and S. Kinne, "Wavelength dependence of the optical depth of biomass burning, urban, and desert dust aerosols," J. Geophys. Res. Atmos., vol. 104, pp. 31333–31349, 1999.

[4] G. P. Gobbi, Y. J. Kaufman, I. Koren, and T. F. Eck, "Classification of aerosol properties derived from AERONET direct sun data," Atmos. Chem. Phys., vol. 7, pp. 453–458, 2007.

[5] J. Lee, J. Kim, C. H. Song, S. B. Kim, Y. Chun, B. J. Sohn, and B. N. Holben, "Characteristics of aerosol types from AERONET sunphotometer measurements," Atmos. Environ., vol. 44, pp. 3110–3117, 2010.

[6] A. H. Omar, D. M. Winker, C. Kittaka, M. A. Vaughan, Z. Liu, Y. Hu, C. R. Trepte, R. R. Rogers, R. A. Ferrare, K.-P. Lee, R. E. Kuehn, and C. A. Hostetler, "The CALIPSO automated aerosol classification and lidar ratio selection algorithm," J. Atmos. Ocean. Technol., vol. 26, pp. 1994–2014, 2009.

[7] H. Che, X. Zhang, X. Xia, P. Goloub, B. Holben, H. Zhao, Y. Wang, X. Zhang, H. Wang, L. Blarel, B. Damiri, R. Zhang, X. Deng, Y. Ma, T. Wang, F. Geng, B. Qi, J. Zhu, J. Yu, Q. Chen, and G. Shi, "Ground-based aerosol climatology of China: aerosol optical depths from the China Aerosol Remote Sensing Network (CARSNET) 2002–2013," Atmos. Chem. Phys., vol. 15, pp. 7619–7652, 2015.

[8] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," Ann. Math. Stat., vol. 27, pp. 832–837, 1956.

[9] X. Wei, Q. Cui, L. Ma, F. Zhang, W. Li, and P. Liu, "Global aerosol-type classification using a new hybrid algorithm and AERONET data," Atmos. Chem. Phys., vol. 24, pp. 5025–5045, 2024.

[10] B. N. Holben, T. F. Eck, I. Slutsker, D. Tanré, J. P. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, and A. Smirnov, "AERONET - a federated instrument network and data archive for aerosol characterization," Remote Sens. Environ., vol. 66, pp. 1–16, 1998.

[11] D. M. Giles, A. Sinyuk, M. G. Sorokin, J. S. Schafer, A. Smirnov, I. Slutsker, T. F. Eck, B. N. Holben, J. R. Lewis, J. R. Campbell, E. J. Welton, S. V. Korkin, and A. I. Lyapustin, "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 measurements," Atmos. Meas. Tech., vol. 12, pp. 169–209, 2019.

[12] O. Dubovik and M. D. King, "A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements," J. Geophys. Res. Atmos., vol. 105, pp. 20673–20696, 2000.

[13] P. Hamill, M. Giordano, C. Ward, D. Giles, and B. Holben, "An AERONET-based aerosol classification using the Mahalanobis distance," Atmos. Environ., vol. 140, pp. 213–233, 2016.

[14] O. Dubovik, B. N. Holben, T. F. Eck, A. Smirnov, Y. J. Kaufman, M. D. King, D. Tanré, and I. Slutsker, "Variability of absorption and optical properties of key aerosol types observed in worldwide locations," J. Atmos. Sci., vol. 59, pp. 590–608, 2002.

[15] S.-K. Shin, M. Tesche, Y. Noh, and D. Müller, "Aerosol-type classification based on AERONET version 3 inversion products," Atmos. Meas. Tech., vol. 12, pp. 3789–3803, 2019.

[16] J. S. Reid, T. F. Eck, S. A. Christopher, R. Koppmann, O. Dubovik, D. P. Eleuterio, B. N. Holben, E. A. Reid, and J. Zhang, "A review of biomass burning emissions Part III: intensive optical properties of biomass burning particles," Atmos. Chem. Phys., vol. 5, pp. 827–849, 2005.

[17] K. N. Liou, An Introduction to Atmospheric Radiation, 2nd ed. San Diego, CA, USA: Academic Press, 2002.

[18] A. Ångström, "The parameters of atmospheric turbidity," Tellus, vol. 16, pp. 64–75, 1964.

[19] C. F. Bohren and D. R. Huffman, Absorption and Scattering of Light by Small Particles. New York, NY, USA: Wiley-Interscience, 1983.

[20] M. C. R. Kalapureddy, D. G. Kaskaoutis, P. Ernest Raj, P. C. S. Devara, H. D. Kambezidis, P. G. Kosmopoulos, and P. T. Nastos, "Identification of aerosol type over the Arabian Sea in the premonsoon season during the Integrated Campaign for Aerosols, Gases and Radiation Budget (ICARB)," J. Geophys. Res. Atmos., vol. 114, D17203, 2009.

[21] D. M. Giles, B. N. Holben, T. F. Eck, A. Sinyuk, A. Smirnov, I. Slutsker, R. R. Dickerson, A. M. Thompson, and J. S. Schafer, "An analysis of AERONET aerosol absorption properties and classifications representative of aerosol source regions," J. Geophys. Res. Atmos., vol. 117, D17203, 2012.

[22] B. W. Silverman, Density Estimation for Statistics and Data Analysis. London, U.K.: Chapman & Hall, 1986.

[23] G. Mie, "Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen," Ann. Phys., vol. 25, pp. 377–445, 1908.

[24] A. S. Mohan, A. Manisekaran, and L. S. Kumar, "Aerosol classification using machine learning algorithms," Indian J. Radio Space Phys., vol. 50, pp. 217–223, 2021.