Article

A Machine Learning Approach to Retrieving Aerosol Optical Depth Using Solar Radiation Measurements

Stavros-Andreas Logothetis [1], Vasileios Salamalikis [2] and Andreas Kazantzidis [1,*]

[1] Laboratory of Atmospheric Physics, Physics Department, University of Patras, GR-26500 Patras, Greece; logothetis_s@ac.upatras.gr

[2] NILU—Norwegian Institute for Air Research, P.O. Box 100, 2027 Kjeller, Norway; vsal@nilu.no

* Correspondence: akaza@upatras.gr

Abstract: Aerosol optical depth (AOD) constitutes a key parameter of aerosols, providing vital information for quantifying the aerosol burden and air quality at global and regional levels. This study demonstrates a machine learning strategy for retrieving AOD under cloud-free conditions based on the synergy of machine learning algorithms (MLAs) and ground-based solar irradiance data. The performance of the proposed methodology was investigated by applying different components of solar irradiance. In particular, the use of direct instead of global irradiance as a model feature led to better performance. The MLA-based AODs were compared to reference AERONET retrievals, which encompassed RMSE values between 0.01 and 0.15, regardless of the underlying climate and aerosol environments. Among the MLAs, artificial neural networks outperformed the other algorithms in terms of RMSE at 54% of the measurement sites. The overall performance of MLA-based AODs against AERONET revealed a high coefficient of determination ($R^2 = 0.97$), MAE of 0.01, and RMSE of 0.02. Compared to satellite (MODIS) and reanalysis (MERRA-2 and CAMSRA) data, the MLA-AOD retrievals revealed the highest accuracy at all stations. The ML-AOD retrievals have the potential to expand and complement the AOD information in non-existing timeframes when solar irradiances are available.

Keywords: aerosol optical depth; machine learning; aerosol retrieval; solar irradiance

1. Introduction

Aerosol particles, suspended in the atmosphere and originating from either natural (e.g., dust and sea salt) or anthropogenic sources (e.g., nitrate and sulfate), exert a direct effect on the Earth's radiative balance through aerosol–radiation interaction and an indirect effect as their role as condensation nuclei in cloud formation [1]. This interaction is contingent upon the aerosol optical properties [2]. While these optical properties are crucial for evaluating the radiative effect of aerosols [3–6], they also hold significance for precise solar resource estimations on both regional and global scales [7].

Reference instruments for measuring solar radiation face limitations in terms of fine temporal resolution and worldwide spatial coverage. Consequently, alternative techniques for solar resource assessment are under scrutiny, with a primary focus on radiative transfer modeling. An accurate assessment of solar resources necessitates a comprehensive understanding of aerosol properties. Integrating these properties into radiative transfer models enables the simulation of long-term solar irradiances under cloud-free conditions [8–17].

There are two primary methods for monitoring atmospheric aerosols. The first method involves the use of ground-based instruments, such as sun photometers. AERONET (Aerosol Robotic NETwork) [18] is a widely recognized ground-based monitoring network of sun photometers that has been providing highly accurate physical and optical characteristics of aerosols at over 600 stations worldwide since 1993. The second method involves utilizing sensors aboard satellites, such as the MODerate resolution Imaging Spectroradiometer (MODIS), Multi-angle Imaging Spectro-Radiometer (MISR), and Cloud-Aerosol

Lidar and Infrared Pathfinder Satellite Observation (CALIPSO). In general, ground-based and satellite-based aerosol retrievals complement each other. Ground-based retrievals excel in accuracy and temporal resolution but face limitations in spatial coverage, particularly in desert areas, which are significant natural sources of suspended aerosol particles in the atmosphere [19–21]. On the other hand, polar-orbiting satellites provide high spatial resolution of aerosol optical depth (AOD) but are limited to offering more than two retrievals per day across a specific region. On the other hand, geostationary satellites, by utilizing their higher temporal resolution, can present almost real-time images and capture the dynamic AOD variations, which is very useful for monitoring events such as forest fires [22]. However, there are two main drawbacks when using geostationary satellites for earth observations. The first is their limited storage and high orbit altitude, which make it challenging to record images with high spatial resolution. The second drawback is that higher uncertainties are observed in AOD retrieval algorithms compared to polar orbit satellites [23]. Considering the spatial and temporal constraints of both satellite and ground-based aerosol retrievals, other techniques have recently emerged, with a predominant focus on AOD.

Various alternative techniques for AOD retrieval reported in the literature fall into four primary categories: (1) backward solving radiative transfer (RT) or clear-sky models using solar radiation measurements [24–27], (2) methodologies based on sunshine duration (SD) measurements [28–30], (3) image processing techniques using sky radiances from all-sky imagers [31–34], and (4) machine learning (ML) and deep learning algorithms employing various independent parameters as input features [35–39].

Salmon et al. [25] retrieved AOD across 16 regions worldwide characterized by different climates and aerosol properties, solving backward two clear-sky radiation models (SMARTS and SOLIS). The model utilized inputs such as broadband direct normal irradiance (DNI), meteorological data, and aerosol information such as single scattering albedo and the Ångström exponent, resulting in an average dispersion error of approximately 0.04 for the retrieved AOD compared to AERONET. Wandji Nyamsi et al. [30] conducted a benchmarking analysis, evaluating three distinct methodologies (linear-based, physical-based, and hybrid) for AOD retrieval based on SD measurements over Europe. The hybrid approach outperformed the other methods, yielding a root mean square error (RMSE) ranging from 0.04 to 0.13 compared to AERONET. Román et al. [32] retrieved AOD at three different wavelengths (465, 536, and 605 nm) using sky radiances from an all-sky imager and the GRASP (Generalized Retrieval of Atmosphere and Surface Properties) [40] code. The predicted AOD agreed well with AERONET retrievals, exhibiting a coefficient of determination of approximately 0.87. Huttunen et al. [36] employed five ML algorithms (MLAs: non-linear regression, artificial neural network, random forest, Gaussian process regression, and support vector machine) to retrieve AOD. All models included global horizontal irradiance (GHI), total column water vapor, and solar geometry as independent parameters. The predicted AOD closely matched observations (absolute bias = 0.04–0.05), with the Gaussian process regression and artificial neural network to outperform other ML algorithms, revealing coefficients of determination of 0.93 and 0.92, respectively.

The primary objective of this study is to implement an ML approach for retrieving AOD, denoted as "MLA-AOD". The innovative aspects of this proposed approach are succinctly outlined below:

- AOD is extracted utilizing the direct beam of solar radiation or direct normal irradiance (DNI) derived from ground-based reference measurements.
- The impact of different solar irradiance components (GHI vs. DNI) on the retrieved AOD is explored, underscoring the significance of selecting the most appropriate solar irradiance component for AOD retrieval.
- The validation process includes comparisons against AERONET measurement locations with diverse climatic and aerosol characteristics.
- MLAs with different prediction mechanisms are benchmarked to identify the "optimal" approach for retrieving AOD.

- The MLA-AOD retrievals are compared with AOD from reanalysis products (MERRA-2 and CAMSRA) through a comprehensive comparative analysis, showcasing the advantages of the proposed methodology at a regional level.

The proposed MLA-AOD methodology presents an aerosol retrieval technique for cloud-free conditions that can be used to reproduce the atmospheric aerosol burden at the high temporal resolution of radiometric instruments. The MLA-AOD approach is evaluated using either pyrheliometer (DNI) or pyranometer (GHI) measurements, presenting different retrieval performances. The selection of the applied solar irradiance component relies on the instrument availability. Even though AERONET sun photometers are established and measured every 5–15 min, it is very common for them to present gaps of hours, days, or weeks or to be inactive for years. This study proposes that the MLA-AOD methodology could be used (a) to fill the AOD gaps in the AERONET sites and (b) to expand the already existing temporal capabilities, regardless of the underlying climate and aerosol conditions. Upgrading the current AOD capabilities is important for the evaluation of satellite products, aerosol data assimilation, air pollution models, AOD trend analyses, climate models, and direct aerosol radiative effect estimation.

The structure of this work is as follows: Sections 2 and 3 present details on the datasets and the proposed methodology. Section 4 covers the results and Section 5 addresses the summary and conclusion of this study.

2. Datasets

2.1. Ground-Based Data

2.1.1. AERONET

AERONET (https://aeronet.gsfc.nasa.gov; accessed on 12 January 2024) comprises a network of sunphotometers, which are ground-based, passive remote sensing instruments [18]. It provides high-quality data on aerosol optical and physical properties, encompassing over 600 stations. AERONET's Direct Sun Algorithm (DSA) derives AOD at seven wavelengths spanning from 340 nm to 1020 nm. DSA relies on spectral solar irradiance measurements directly on the solar disk, captured from the CIMEL Electronique CE318 multiwavelength sunphotometer. Besides AOD, DSAA also provides information about total column water vapor (TCWV), based on measurements at 940 nm. AERONET, except for DSA, applies an inverse algorithm using sky radiances in the almucantar geometry to provide other aerosol properties like the single scattering albedo, aerosol volume distribution, phase function, refractive index, etc. [41].

In this study, AOD at 550 nm (AOD_{550nm}) is utilized to maintain consistency with MODIS, CAMSRA, and MERRA-2 reanalysis. The AERONET AOD is interpolated at 550 nm using the Ångström law. For subsequent analyses, cloud-screened data rounded to the nearest minute are employed, meeting quality criteria established by pre-field and post-field calibrations (Level 2.0 data from Version 3; L2V3) [42]. The spectral AOD uncertainties in the visible spectrum for AERONET Version 3 are reported to be 0.01 [42].

2.1.2. BSRN

The Baseline Surface Radiation Network (BSRN, https://bsrn.awi.de; accessed on 12 January 2024), is a ground-based solar radiation monitoring network that has been measuring solar radiation at various locations worldwide since 1992 [43,44]. BSRN provides comprehensive data on the various components of incoming solar radiation, namely global horizontal irradiance (GHI), direct normal irradiance (DNI), and direct horizontal irradiance (DHI). GHI refers to the total amount of shortwave radiation reaching the ground surface. DNI is the amount of solar radiation received per unit area by a surface that is always normal to the solar rays. DHI is the total amount of solar radiation that reaches the ground and does not arrive on a direct path from the sun. The network delivers quality-assured measurements at a 1 min temporal resolution. Each BSRN radiometric station employs a first-class pyranometer and a pyrheliometer to measure GHI and DNI, respectively. The raw observations from BSRN undergo quality control [45] using the procedures outlined in

the SolarData R package [46]. For subsequent analysis, only the 1 min quality-assured DNI and GHI measurements from BSRN are utilized.

2.1.3. Stations' Characteristics

Figure 1 displays the spatial distribution of the collocated AERONET-BSRN stations (N = 26), while Table S1 provides information about the station names and data availability. The majority of the stations are situated in Europe and the US (Figure 1). The five distinct colors in Figure 1 represent the Köppen–Geiger (KG) [47] climate clusters (CCs). The KG climate classification system categorizes the world's climate into five primary CCs based on air temperature and precipitation amount, namely (A) Equatorial, (B) Arid, (C) Warm Temperate, (D) Snow, and (E) Polar. The corresponding climate zone for each station is determined using high-resolution data (100 arc sec) via the 'kgc' R package [48]. Each CC is represented by at least one station, with arid and temperate climates being the most prevalent in terms of station occurrence.
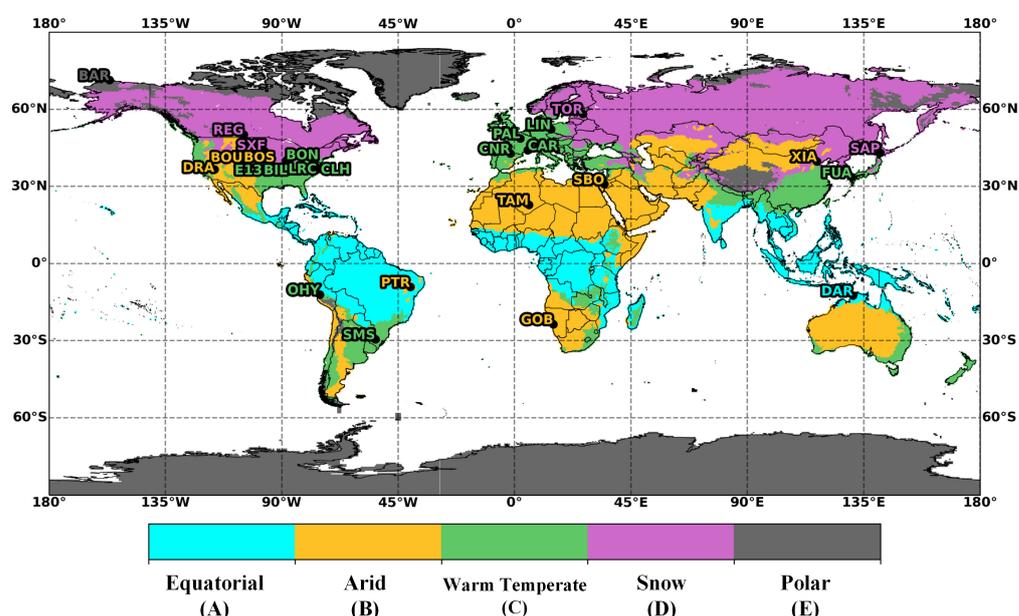

Figure 1. Spatial distribution of the AERONET-BSRN stations. The five different colours correspond to the five climate classes based on KG climate classification.

It is significant to evaluate the proposed methodology in diverse environments to assess its resilience to potential changes in climate and aerosol types. Previous studies showed that the accuracy of model-based and reanalysis AOD against observations varied climatically [49,50] and regionally [50,51]. For instance, the training of an ML algorithm to retrieve AOD becomes more challenging in regions potentially affected by dust intrusions or wildfires.

2.2. Reanalysis Data

2.2.1. CAMSRA

The Copernicus Atmosphere Monitoring Service Reanalysis (CAMSRA) represents the latest global dataset for atmospheric composition, created by the European Centre for Medium-Range Weather Forecasts (ECMWF). This dataset consists of 3-D time-consistent fields of atmospheric composition, incorporating aerosols and various chemical species [52]. CAMSRA was generated through 4DVar data assimilation within CY42R1 of ECMWF's Integrated Forecast System (IFS), in which meteorological modeling is incorporated. The 4DVar analysis applies IFS dynamics and physics to produce a sequence of atmospheric states that closely align with the available observations.

Concerning AOD_{550nm}, CAMSRA assimilates satellite-derived AOD_{550nm} retrievals obtained from MODIS Collection 6 aboard Aqua and Terra, as well as the Advanced Along-Track Scanning Radiometer (AATSR) aboard Envisat [53]. For TCWV, CAMSRA assimilates satellite-based infrared and microwave radiances from various sensors onboard satellites such as AMSR-2 (GCOM-W1), AMSRE (AQUA), GMI (GPM), and TMI (TRRM), among others. Since 2003, CAMSRA data have been provided at approximately 80 km resolution and on a 3 h basis. In this study, AOD_{550nm} and TCWV retrievals are acquired through CAMSRA. The validation of CAMSRA AOD_{550nm} against AERONET observations revealed a global Root Mean Square Error (RMSE) of 0.144 [50]. The validation of CAMSRA TCWV against AERONET retrievals revealed satisfactory performance with a coefficient of determination (R^2) of 0.93 and an RMSE of 0.3 cm (Figure S1).

2.2.2. MERRA-2

The Modern-Era Retrospective Analysis for Research and Applications-2 (MERRA-2) is the latest global reanalysis dataset developed by the NASA Global Modeling and Assimilation Office (GMAO). MERRA-2 assimilates meteorological and aerosol measurements obtained from a variety of ground-based and satellite-based products [54–56]. This analysis focuses on AOD_{550nm} for intercomparison purposes. The MERRA-2 AOD_{550nm} reanalysis product incorporates AOD products spanning different time periods. More specifically, before 2000, AOD_{550nm} relies exclusively on AVHRR retrievals. Subsequently, MERRA-2 assimilates data from MODIS (onboard both Terra and Aqua, covering the period from 2000 and 2002 to the present, respectively), MISR (from 2000 to 2014), and AERONET (from 1999 to 2014). MERRA-2 data (product name: tavg1_2d_aer_Nx) are provided at a spatial resolution of $0.5° \times 0.625°$ on an hourly basis.

The global validation of MERRA-2 AOD_{550nm} against AERONET data yielded an RMSE of 0.126 [50], outperforming CAMSRA. The enhanced performance of MERRA-2 AOD_{550nm} is attributed to its utilization of AERONET retrievals as part of its data assimilation methodology.

2.3. Satellite Data

MODIS

MODerate resolution Imaging Spectroradiometer (MODIS) is a passive sensor that has been hosted on NASA's Terra and Aqua polar satellites since 2000 and 2002, respectively. MODIS monitors several pivotal atmospheric constituents, like aerosols [57], on a global scale. MODIS takes advantage of its wide swath (~2330 km) to provide approximate global coverage by making 14–15 orbits daily. The Aqua/Terra crosses the equator in a northbound/southbound direction at 13:30/10:30 UTC. Across dark (in the visible spectrum) surfaces over land [58–60] or/and ocean [61–63] areas, the Deep Target algorithm (DT) is applied, while over bright-reflecting land surfaces [64,65], including desert, semi-arid, and urban regions, the Deep Blue algorithm (DB) is adopted.

In this study, the AOD_550_Dark_Target_Deep_Blue_Combined product from Collection 6.1 from Aqua and Terra is applied, which is a merged dataset of DB and DT algorithms. In particular, the AOD values are merged when specific quality assurance (QA) conditions are met (QA $\geq$ 1 for DT ocean, QA = 3 for DT land, and QA $\geq$ 2 for DB land).

3. Methodology

The proposed methodology (MLA-AOD) for retrieving AOD_{550nm} at 1 min resolution is illustrated in Figure 2. Various data preprocessing steps are included to increase the performance of the MLAs.
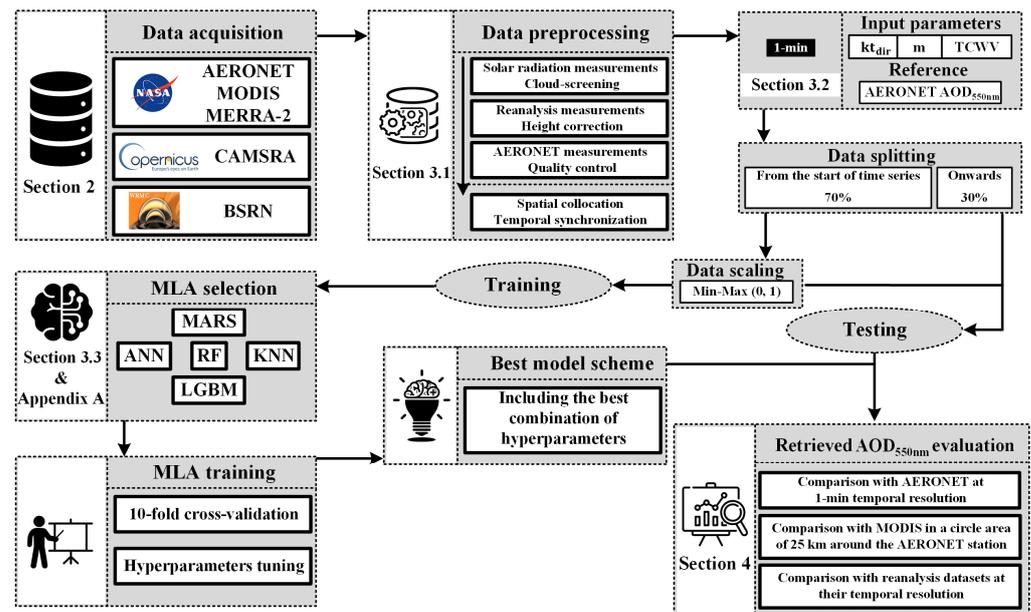
Figure 2. Flowchart of the MLA-AOD retrieval methodology.

The applied methodology involves two primary steps: data acquisition and preprocessing, which are extensively detailed in Sections 2 and 3.1. Subsequently, the preprocessed dataset is divided into two subsets, namely the training and testing datasets. The splitting strategy follows a 70/30 approach without using a random-sampling approach. The first 70% of the time series represents the training set and the remainder represents the test set. Table S1 includes information about the training and testing periods for each station.

The independent (input) features undergo normalization between 0 and 1 using the Min–Max normalization method. Five different ML algorithms (see Section 3.3 and Appendix A) are evaluated to retrieve AOD. The MLA structure is determined through several internal parameters called hyperparameters. The optimal configuration for these hyperparameters for each MLA is determined through a randomized search procedure during training. The full cartesian hyperparameter space is given in Table S2. The randomized search involves a 10-fold cross-validation process, with MSE serving as the fitness function. For each random hyperparameter set, the data are randomly divided into 10 roughly equal subsamples. Then, at each iteration, one of those 10 subsamples is retained as the "validation" dataset for MLA evaluation, and the 10-1 subsamples train the ML model. The MSE scores retrieved through the 10 iterations are averaged to retrieve a total score which is characteristic for each hyperparameter set. This process is repeated 100 times. The "optimal" hyperparameter combination for each MLA is determined based on the minimum of total scores. The MLA-AOD methodology (Figure 2) is applied for each collocated AERONET-BSRN station (see Section 2.1.3).

3.1. Data Preprocessing

The raw data obtained from ground-based measurements are cloud-screened to isolate clear sky conditions. Although L2V3 AERONET retrievals are described as cloud-screened, the temporal merging of AERONET and BSRN data may introduce cloud-contaminated cases, attributed to the smaller aperture of sun photometers compared to pyrheliometers [25]. To ensure clear-sky conditions, two distinct filters were applied—one for DNI and one for GHI. Concerning DNI, within one hour around each BSRN measurement, the standard deviation is calculated. If it is above 20% of the corresponding BSRN DNI measurement, the data point is detected as cloudy and removed [25]. In addition, even though the sun's disk could be cloudless, the rest of the sky can contain sparse clouds that affect the GHI measurements. Therefore, to avoid possible cloud contamination, the Reno and Hansen clear-sky detection methodology [66] is applied to the GHI measurements.

This method is preferred here over other published methodologies because it relies on periods of time. In particular, the Reno and Hansen clear-sky detection methodology [66] is applied on time windows (here, 10 min) of 1 min temporal resolution. If at least one minute is assigned as cloudy, the whole window is flagged as cloudy and eliminated from the analysis. By applying the filters for GHI and DNI described above, the cloudy solar irradiance measurements are unlikely to be found in the used data. Subsequently, the cloudless solar irradiance measurements are synchronized with AERONET 1 min retrievals.

Moreover, the reanalysis data (TCWV and AOD_{550nm}) are extracted at the ground-based stations through nearest-neighbor interpolation. The 3-hourly TCWV from CAMSRA is linearly interpolated at 1 min temporal resolution, aligning with the AERONET and BSRN datasets.

For validation purposes, both MLA-AOD based and AERONET AODs are temporally aggregated to match the resolution of the reanalysis and satellite products. Given that MERRA-2 hourly data are available at half-hour intervals (e.g., 00:30, 01:30, etc.), the ground observations are averaged over ± 30 min intervals, centered on the MERRA-2 timestamps. On the other hand, CAMSRA provides data with a 3 h resolution (e.g., 00:00, 03:00, etc.). In this case, the ground-based data are represented as 90 min averages, centered on the CAMSRA 3 h intervals. Regarding the satellite data, the MODIS AODs are spatially averaged if at least one measurement resides within a circle area of 25 km around the AERONET station. Then, the spatially averaged AODs are temporally averaged when at least one of the AERONET measurements falls within a time window of 2 h around the satellite overpass time.

Another significant point in validation studies involves addressing elevation differences between ground-based and reanalysis datasets [50]. To mitigate these discrepancies, the reanalysis AOD is adjusted to the station elevation using the scale-height correction of Gueymard and Yang [50].

In addition, any unrealistic AERONET AOD retrievals ($AOD_{550nm} < 0$) are eliminated for the subsequent analysis. Finally, the proposed methodology is applied using data for a 15-year period (during 2004–2019). The data period is related to the station's data availability. Finally, for validation against reanalysis, only stations with at least a minimum of 100 (3 h) CAMSRA and 300 (1 h) MERRA-2 records are retained.

3.2. Input Parameters

The direct and global clearness indexes are calculated using Equations (1) and (2), respectively,

$$kt_{dir} = \frac{DNI}{E_0} \tag{1}$$

$$kt_{glob} = \frac{GHI}{E_0} \tag{2}$$

where E_0 is the received irradiance on a horizontal plane at the top of the atmosphere. The optical air mass is calculated based on Kasten and Young's [67] formula, using Equation (3),

$$m = (P/P_0) / \left[\cos(SZA) + 0.50572 \cdot (SZA + 6.07995)^{-1.6364} \right] \tag{3}$$

where P and P_0 are the pressure level at the station's elevation and sea level, respectively, while SZA corresponds to the solar zenith angle. The TCWV in cm is derived through the CAMSRA (see Section 2.2.1).

The retrieval of AOD through remote sensing always requires information about the solar geometry and amount of the non-aerosol atmospheric constituents. For example, the AERONET retrieval methodology relies on Beer–Lambert–Bouguer law, considering the impact of solar geometry and TCWV [18,42]. In addition, solar geometry is considered when the actual and estimated reflectances are compared through AOD retrieval from satellite measurements [68]. In the proposed methodology, two different solar irradiance

components were applied, presenting different levels of alteration with variations in aerosol burden, solar geometry, and TCWV.

Figure 3 depicts the influence of AOD_{550nm} on DNI and GHI under various SZA (Figure 3a) and TCWV (Figure 3b) cases. The solar irradiance simulations were performed via the RTM LibRadtran package [69,70]. DNI is apparent more sensitive than GHI as AOD_{550nm} increases, which is consistent with the study conducted by Ruiz-Arias and Gueymard [71], becoming a very influential parameter for feeding the MLAs. Setting a constant water content (Figure 3a), the solar irradiance drops substantially as SZA increases. For instance, for SZA of 10° (70°), the DNI decreases from ~1000 Wm^{-2} (~300 Wm^{-2}) to ~350 Wm^{-2} (~10 Wm^{-2}) at higher AOD_{550nm} (Figure 3a). Therefore, it is essential to include information on solar geometry in MLAs. For example, if a DNI of 600 Wm^{-2} is considered (Figure 3a, under constant TCWL level), this value can be simulated from the RTM from various AOD values varying from 0.02 (for SZA = 50°) to 0.80 (for SZA = 10°). Based on Figure 3b, TCWV affects DNI to a lesser extent than SZA. For instance, if DNI equals 600 Wm^{-2}, without knowing TCWV, the AOD_{550nm} ranges from 0.4 (for TCWV = 4 cm) to 0.75 (for TCWV = 0 cm). Therefore, clear-sky solar irradiance is predominantly influenced by solar geometry, aerosols, and water vapor. Therefore, all three variables jointly provide essential information for retrieving AOD_{550nm}.
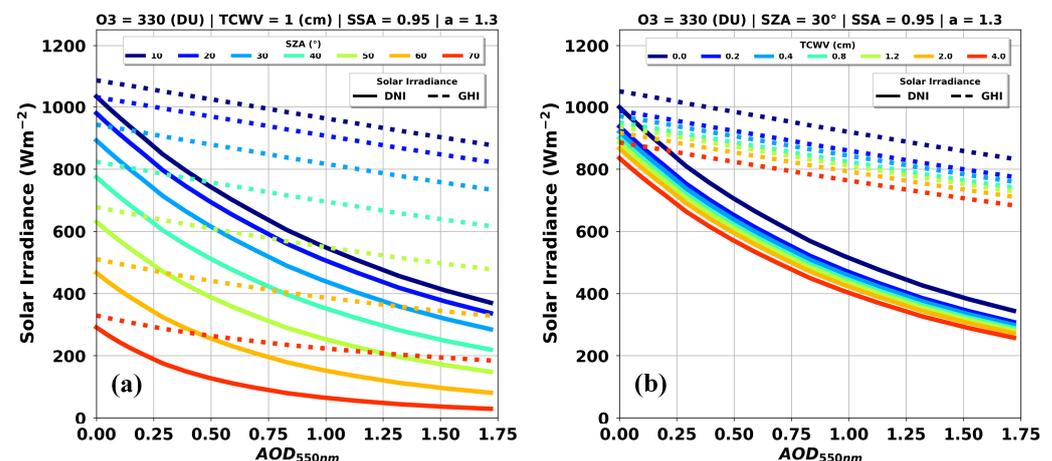

Figure 3. Global horizontal (GHI) and direct normal (DNI) irradiance vs. AOD_{550nm} for different (**a**) solar zenith angle (SZA) and (**b**) total column water vapor (TCWV) cases. The solid and dashed lines correspond to GHI and DNI. In the radiative transfer simulations, total column ozone (O_3), single scattering albedo (SSA), and the Ångström exponent (**a**) are taken as equal to 330 D.U., 0.95, and 1.3, respectively.

3.3. Machine Learning Algorithms

In this study, five machine learning models with different prediction mechanisms, including linear, distance-based, and tree-based approaches, are utilized. This setup establishes a benchmarking task aimed at evaluating and comparing the effectiveness of these ML algorithms in simulating AOD. The five MLAs used in this study are Light Gradient Boosting Machine (LGBM: tree-based), Random Forest (RF: tree-based), Multivariate Adaptive Regression Splines (MARS: linear-based), K-Nearest Neighbors (KNN: distance based), and Artificial Neural Network (ANN: linear based). For the sake of brevity, a comprehensive description of each MLA is presented in Appendix A.

4. Results

4.1. Direct Normal Irradiance vs. Global Horizontal Irradiance as Model Input

This sub-section investigates the performance of the applied methodology using various ML models in terms of the direct and global clearness index. Two distinct scenarios are examined: Scenario 1 considers the three input parameters (see Section 3.2), using

the kt_{dir} as a solar irradiance component. Conversely, in Scenario 2, kt_{dir} is substituted by kt_{glob}. Figure 4 depicts the mean absolute error (MAE) and the relative MAE between Scenario 1 and 2, using solely the retrievals of the "optimal" MLA, compared to AERONET measurements. More specifically, the "optimal" MLA is jointly assessed using the MAE, root mean square error (RMSE), and correlation coefficient (R) (Figure S2), assuming an equal contribution for each statistical metric. The "best" MLA at each station is presented in Table S1. The linear correlation between MLA-AOD and AERONET individually for each station per KG climate class is depicted in Figures S3–S7.
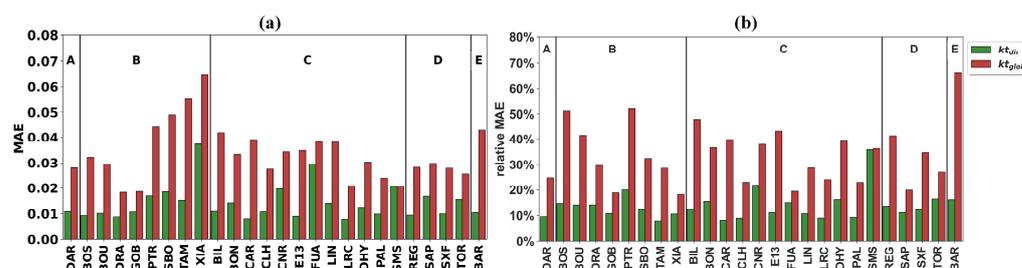
Figure 4. (**a**) Mean absolute error (MAE) and (**b**) relative MAE between MLA-AOD and AERONET AOD for Scenario 1 (kt_{dir}) and Scenario 2 (kt_{glob}). For the MLA-AOD, the "optimal" algorithm is applied. The horizontal lines separate the stations based on the KG classes (A = Equatorial, B = Arid, C = Warm Temperate, D = Snow, and E = Polar).

The selection of solar irradiance type as an input feature significantly affects the model performance. Scenario 1 reveals the optimal model performance as depicted in Figure 4. The MAE (relative MAE) values ranges are 0.007–0.037 (7.9–35.9%) and 0.018–0.64 (18.1–66.3%) for Scenario 1 and 2, respectively, with the corresponding metrics for each station and Scenario 2 being approximately twice as high. In terms of R, Scenario 1 consistently outperforms Scenario 2, extending from 0.50 to 0.99 (Figure S2). Both scenarios show that the highest RMSEs are recorded for XIA. This can be attributed to the high aerosol variability (Figure S4) in that area due to the high air pollution levels due to local fossil fuel burning from automobiles, the transported polluted plumes from nearby regions, and the transport of dust particles [72]. Despite the relatively high RMSE values in XIA, the relative errors are lower than 10%. Similar results were observed for other stations with high AOD variability such as TAM and SBO (Figure S4).

The difference in the observed performance between the two scenarios is primarily attributed to the impact of aerosol load on GHI and DNI. As shown in Section 3.2 and previous studies [71,73], DNI is highly sensitive to aerosol burden improving the performance of MLAs when it is used as an input parameter.

The performance among MLAs is quite similar at each station, highlighting the model's versatility (Figure 5 and Figure S2). Nevertheless, for both scenarios, ANN shows the highest number of appearances (Figure S2). For Scenario 1, the majority of the stations had the lowest MAE (Figure 5) values for ANN (12/26), followed by MARS (6/26) and LGBM (5/26). Regarding RMSE (Figure S2c), again, most of the stations recorded their lowest values for ANN (14/26), followed by MARS (8/26) and RF (4/26). Similar results are revealed for the stations in Scenario 2, with ANN being the MLA with the lowest values either for MAE or RMSE (Figure S2). Altogether, ANN and MARS are the most frequently "optimal" MLAs with the most appearances (22 out of 26 of the stations). The following sections consider the input parameters from Scenario 1 (the direct beam solar irradiance showed greater performance against GHI) and the "optimal" MLA.
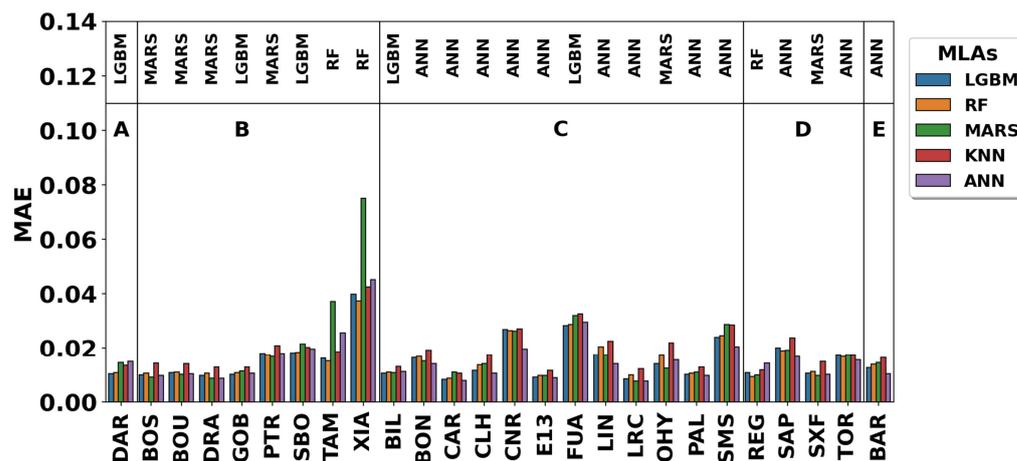
Figure 5. Mean absolute error (MAE) for Scenario 1 for each machine learning algorithm (MLA). The horizontal lines separate the stations based on the KG classes (A = Equatorial, B = Arid, C = Warm Temperate, D = Snow, and E = Polar). The MLA with the best performance is added above each station.

An interesting finding of the presented methodology is the adequate performance of MLAs at various climates, regions, and aerosol types (Figures 4 and S2). For example, at Tamanrasset (TAM) station, located in Southern Algeria, the main aerosol composition consists of mineral dust particles originating from the Sahara Desert, with a relatively high average AOD over time (here, AOD_{550nm} = 0.17) [74,75]. Those aerosol particles are quite absorbing. Although the applied methodology is blind regarding the aerosol size or absorptivity, it can retrieve AOD_{550nm} with reasonable accuracy (Scenario 1; MAE = 0.015–0.037, R = 0.95–0.99). On the other hand, Carpentras, France (CAR), is placed near industrial and urban areas with fine-mode particles as the primary aerosol size. According to Logothetis et al. [75], CAR is mainly affected by fine-absorbing (black carbon) and non-absorbing (sulfate and nitrate) particles. The applied methodology proved to be adequate for accurately retrieving AOD_{550nm} (Scenario 1; MAE = 0.007–0.011, R = 0.97–0.98) over CAR.

MLA-AOD retrieval performance is also investigated for stations with mixed aerosol conditions, including both fine and coarse aerosol modes. For example, Sede Boker (SBO), Israel, is frequently affected by mineral dust particles emitted by nearby desert areas and fine-mode absorbing aerosols from local sources [75]. The relevant errors in AOD_{550nm} range from 0.018 to 0.021 (Figure 5) and the correlation coefficient ranges from 0.94 to 0.96 for the various MLAs in Scenario 1 (Figure S2).

The overall performance of MLA-AOD retrievals, including the "optimal" MLA and Scenario 1, is investigated against AERONET in Figure 6. MLAs document a high coefficient of determination (R^2 = 0.97), MAE of 0.01, and RMSE of 0.02 (Figure 6), highlighting the adequate performance of MLA-AOD methodology compared to AERONET. The majority of the AOD values were lower than 0.5, with the MLA-AOD performance being adequate (see the second plot in Figure 6). For high AOD values (>0.5), the MLA-AOD retrievals have the tendency to underestimate the AERONET measurements.

4.2. The Effect of Aerosol Properties on MLAs Retrieval Performance

The relationship between the differences of the MLA-AOD and observed AOD and the AOD magnitude is investigated (Figure 7). More specifically, the error is drawn against various AOD bins based on AERONET observations (Figure 1).

Based on Figure 7a, the ML-based retrievals encompass higher deviations at higher AOD_{550nm}. For AOD_{550nm} lower than 0.4, the mean differences are lower than 0.02 regardless of the underlying climate conditions. The bin frequency at each CC for AOD_{550nm} < 0.4 is relatively high, including at least 1000 observations. Within high AOD_{550nm} ranges (>0.4), the deviations of ML estimations are relatively low (between ±0.05), highlighting

the model's high accuracy for extreme AOD_{550nm} values. In addition, for high AOD_{550nm} ranges), the MLA-AOD revealed the tendency to underestimate the AERONET AOD. As expected, the frequency of AOD_{550nm} at high aerosol classes is low, since high AOD_{550nm} conditions are found in or close to deserts and heavy pollution areas. Additionally, the station density is low in the equatorial and polar climate clusters.
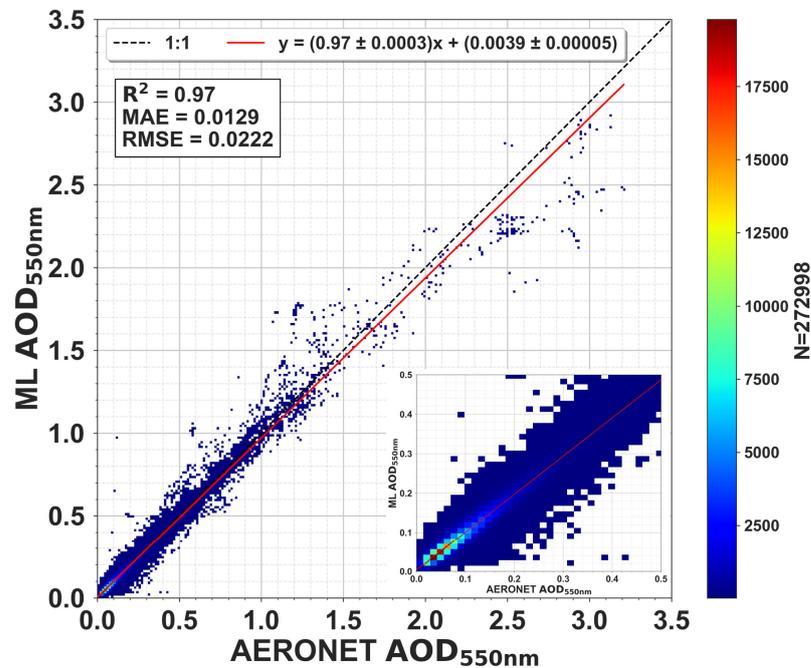

Figure 6. Density scatter plots of MLA-AOD_{550nm} AERONET AOD_{550nm} retrievals. The black dashed line is the identical line and the red solid line corresponds to the linear regression fit. On each plot, a second plot is presented, zoomed in for AOD values ranging between 0 and 0.5.
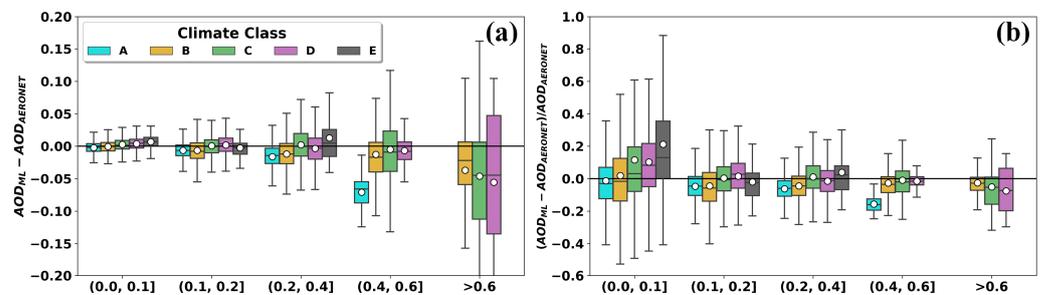
Figure 7. (**a**) Differences and (**b**) relative differences between ML-AOD and AERONET at specific AOD_{550nm} bins. The colors at each station correspond to the KG climate classes (A = Equatorial, B = Arid, C = Warm Temperate, D = Snow, and E = Polar).

The relative differences, using the AERONET AOD as a skill reference, are calculated to investigate the relationship between the AOD_{550nm} deviations and their magnitudes (Figure 7b). For all AOD_{550nm} cases, the relative differences are within 0.20 (Figure 7b). The relatively high documented variability at high AOD_{550nm} bins (Figure 7a) is reduced when relative differences are used (Figure 7b).

Based on the discussion above, MLAs seem to accurately reproduce the observed AOD in regions with different aerosol characteristics, even if information about the type, size, and absorptivity is absent. The dependency of the MLA-AOD retrieval performance on aerosol properties indicators of aerosol size ($a_{440-870nm}$) and absorptivity (SSA_{440nm}) is investigated (Figure 8) by calculating the error between the MLA-AOD and AERONET AOD within specific $a_{440-870nm}$ and SSA_{440nm} bins based on AERONET observations. It is worth mentioning that aerosol properties are not incorporated as independent parameters

in the MLA-AOD, since the data availability for SSA is limited in the AERONET sites, and also to keep the MLAs as simple as possible. Regarding aerosol size (Figure 8a), the mean differences are lower than 0.02 regardless of the $a_{440-870nm}$ bin. Relatively higher variability and mean error are observed for coarse ($a_{440-870nm} < 0.4$) and fine particles ($a_{440-870nm} > 2.0$), respectively. Regarding aerosol absorptivity (Figure 8b), the mean differences are also lower than 0.02 regardless of the SSA_{440nm} values. A relatively higher mean error is observed for highly absorbing aerosols ($SSA_{440nm} \leq 0.85$).
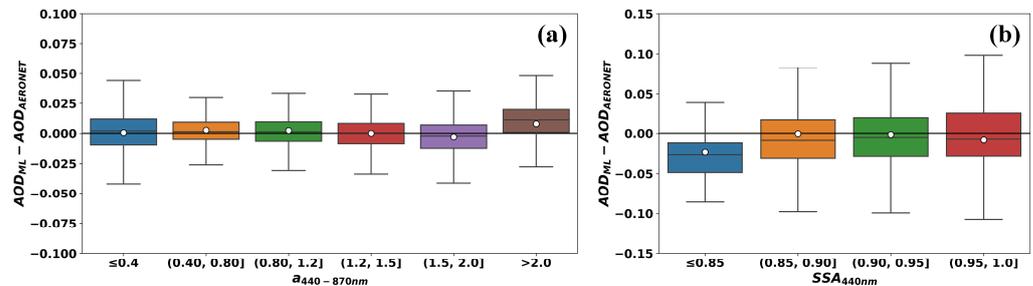
Figure 8. Differences between ML-AOD and AERONET AOD at specific (**a**) Ångström exponent between 440 and 870 nm ($a_{440-870nm}$) and (**b**) single scattering albedo at 440 nm (SSA_{440nm}). The number of AERONET $a_{440-870nm}$ and SSA_{440nm} retrievals were equivalent to 272,998 and 450, respectively.

Nevertheless, the incorporation of secondary parameters (e.g., single scattering albedo) regarding aerosols into the MLA design could increase the performance of MLA. However, the limited availability of this type of aerosol information supports the decision to evaluate simple MLA with easily found independent predictors such as GHI or DNI, TCWV, and SZA.

4.3. MLA-AOD vs. MODIS

In this section, a comparative analysis between the MLA-AOD and MODIS AOD (both Terra and Aqua) is performed. As described in Section 4.1, only the "best" MLA for each station is used for intercomparison purposes. Figure 9 depicts the RMSE values for the MODIS and MLA-AOD (for Scenario 1, see Section 4.1) against AERONET data. The MLA-AOD outperform MODIS at nearly all stations (except BAR). MODIS RMSEs vary between 0.03 (BAR) and 0.18 (SBO) (Figure 9). Regarding MLA-AOD, RMSE ranges from 0.01 (CAR) to 0.06 (XIA) for the MLA AOD_{550nm}.
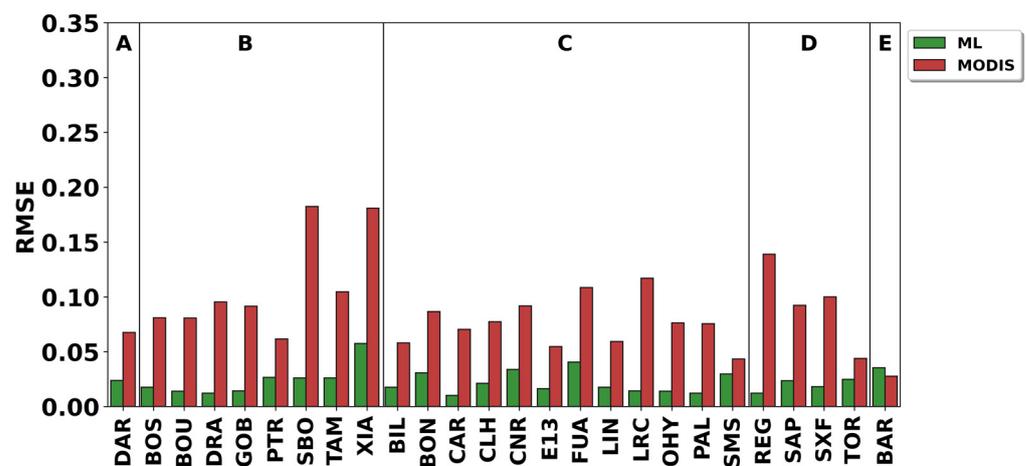

Figure 9. Root mean square error values for ML-AOD and MODIS AOD_{550nm} against AERONET retrievals. The horizontal lines separate the stations based on the KG classes (A = Equatorial, B = Arid, C = Warm Temperate, D = Snow, and E = Polar).

The MLA-AOD and MODIS are linearly compared against AERONET in Figure 10. MLA-AOD documents a high coefficient of determination (R^2 = 0.98) and RMSE of 0.02 (Figure 10), while the MODIS Aqua and Terra datasets reveal lower R^2 (0.70) and higher RMSE (0.1) values (Figure 10). The MAE values were 0.02 and 0.07 for MLA-AOD and MODIS, respectively. The comparison between the proposed methodology and MODIS AOD revealed the adequate performance of the MLA-AOD methodology.
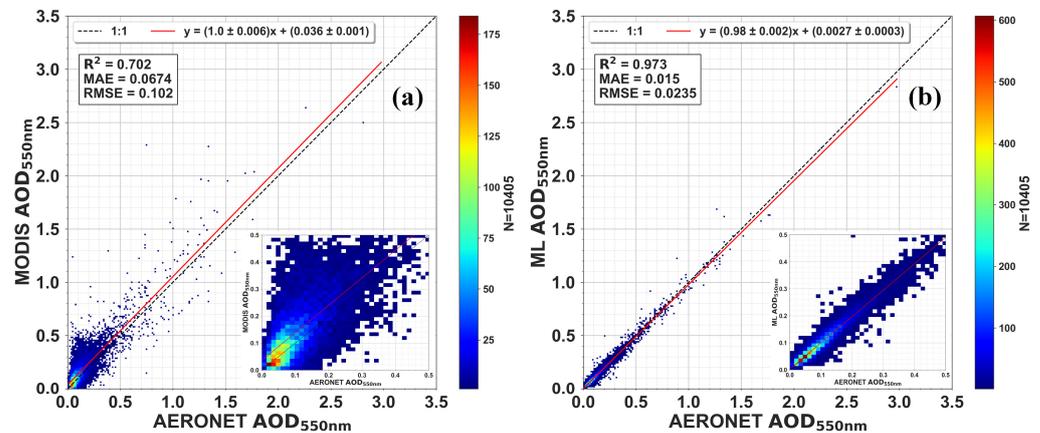

Figure 10. Density scatter plots of AERONET AOD_{550nm} retrievals against (**a**) MODIS, and (**b**) MLA-AOD. The black dashed line is the identical line and the red solid line corresponds to the linear regression fit. On each plot, a second plot is presented, which is zoomed in for AOD values ranging between 0 and 0.5.

Figure S8 presents the daily variability of AERONET, ML-AOD, and MODIS AOD_{550nm} for specific months and stations. For the TAM station (Figure S8a), all three AOD products have the same pattern for all days in May 2015. The MODIS performance in this case is adequate. Nevertheless, the MLA-AOD provides very accurate predictions, which are almost identical to AERONET measurements. For the XIA station (Figure S8b), the AOD variability in March 2015 was between 0.2 and 1.3. Again, all products follow the daily AOD pattern, but retrievals from the proposed methodology are more accurate, with MODIS having the tendency to overestimate the AERONET measurements. Similar results are observed at the other stations (Figure S8c–f), revealing that the MLA-AOD methodology captures the daily variability of AERONET measurements against MODIS with higher accuracy.

4.4. MLA-AOD vs. Reanalysis Products

Globally, the primary source of aerosol properties originates from spaceborne remote sensing instruments (e.g., MODIS in the previous section). However, aerosol properties can be obtained through aerosol reanalysis datasets at a good temporal resolution. AOD_{550nm} from MLAs is compared against MERRA-2 and CAMSRA products. As shown in Section 4.1, the "best" MLA for each station is selected for intercomparison purposes.

Figure 11 illustrates the results obtained for each reanalysis dataset and the applied methodology (for Scenario 1, see Section 4.1 against AERONET data. The MLA-AOD outperform MERRA and CAMSRA at all stations and climate types. Regarding MERRA-2, RMSE varies between 0.03 (SMS) and 0.3 (XIA) (orange bars in Figure 11). Accordingly, RMSE ranges from 0.01 (CAR) to 0.1 (XIA) for the MLA AOD_{550} (blue bars in Figure 11). CAMSRA RMSE values lie within 0.04 (SMS) and 0.24 (XIA), providing higher deviations than MERRA-2 in 23 out of 27 stations. It is worth mentioning that MLA-AOD follows the error patterns of the reanalysis dataset but with significantly lower magnitudes. In the majority of the stations, the MLA-AOD approach reveals RMSE values 2–3 times lower than reanalysis datasets compared to AERONET retrievals.
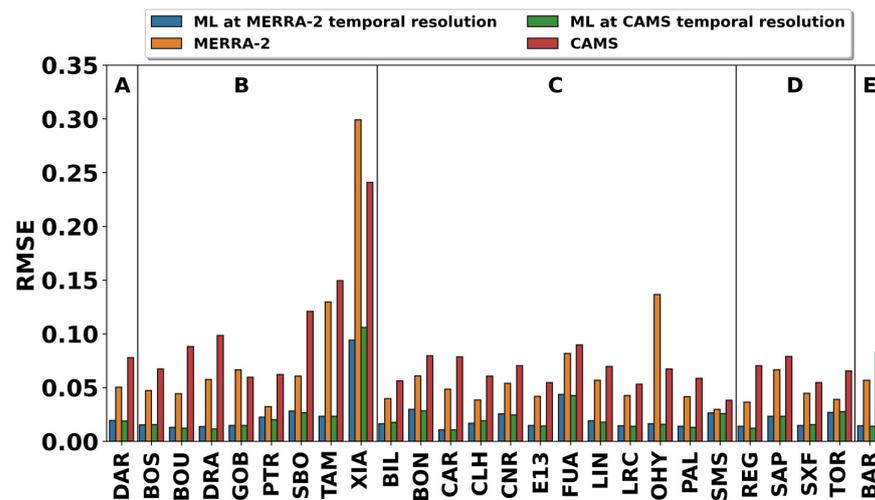
Figure 11. Root mean square errors for ML-AOD against CAMSRA and MERRA-2 AOD$_{550nm}$. The horizontal lines separate the stations based on the KG classes (A = Equatorial, B = Arid, C = Warm Temperate, D = Snow, and E = Polar).

The former findings are consistent with Gueymard and Yang's [50] study. As they reported, the greater performance of MERRA-2 against CAMSRA can be attributed to the different assimilated remote sensing datasets. Apart from the data derived from spaceborne instruments, MERRA-2 also assimilates the AERONET measurements during 1999–2014. The reanalysis and MLA-AOD are compared against AERONET in Figure 12. All stations are used independently based on their climate type. MLAs document a high coefficient of determination ($R^2 = 0.97$) and RMSE of 0.02 (Figure 12), while the reanalysis datasets reveal lower R^2 (MERRA-2 = 0.72, CAMSRA = 0.70) and higher RMSE (MERRA-2 = 0.07, CAMSRA = 0.09) values (Figure 12).
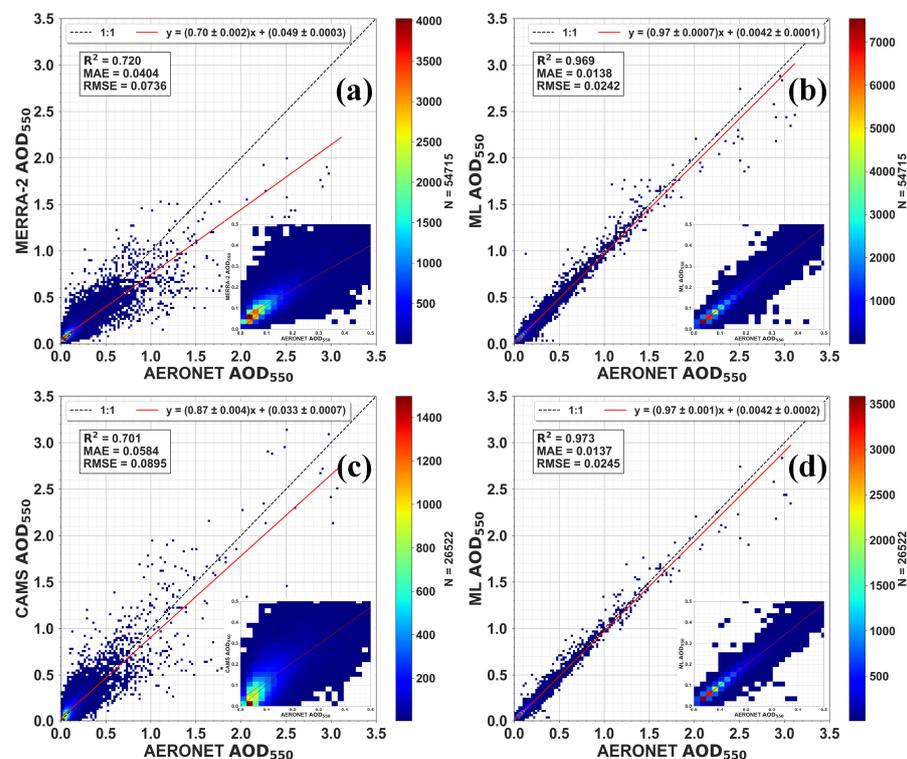

Figure 12. Density scatter plots of AERONET AOD$_{550nm}$ retrievals against (**a**) MERRA-2, (**b**) MLA-AOD at MERRA-2 temporal resolution, (**c**) CAMS, (**d**) MLA at CAMS temporal resolution AOD$_{550nm}$.

The black dashed line is the identical line and the red solid line corresponds to the linear regression fit. On each plot, a second plot is presented, zoomed in for AOD values ranging between 0 and 0.5.

Consequently, the presented findings verify that MLAs have the potential to retrieve AOD from DNI, being unaware of any other aerosol characteristics during the observation period, resulting in more accurate data than those provided by CAMSRA or MERRA-2.

5. Summary and Conclusions

In this study, a machine learning-based approach (MLA-AOD) is evaluated to estimate AOD at 550 nm using solar irradiance data from ground-based observations, water vapor column (taken from the CAMS reanalysis dataset) and solar geometry. The data are imported into five MLAs with different theoretical prediction mechanisms.

Sensitivity analysis between GHI and DNI is conducted to interpret which solar component (GHI and DNI) is the most appropriate for AOD predictions. The algorithms' performance provides errors up to two times lower when DNI is used against GHI. Among the MLAs, the ANN provided the highest retrieval performance at most stations. The MLA-AOD methodology performance is also investigated under various climates. For low (<0.4) AOD values, all climate classes showed errors around 0.02, while for higher AOD records, the respective biases slightly increased, tending to underestimate the AERONET AOD. Overall, the MLA-AOD methodology documented a high R^2 = 0.97, MAE of 0.01, and RMSE of 0.02 compared to AERONET. In addition, MLA-AOD retrievals were accurate in regions with different aerosol optical characteristics, even if information about the type, size, and absorptivity was absent during the model training. The proposed methodology was compared to MODIS AOD, outperforming the satellite data at nearly all stations. The MLA-AOD retrievals were also compared to MERRA-2 and CAMSRA reanalysis products. At all sites, the MLA-AOD outperformed reanalysis data, reaching R^2 = 0.97 and an RMSE = 0.02.

Considering the great performance of MLAs when reproducing AOD, the results of the proposed methodology could be used to fill and extend either the existing or missing AOD time series derived from ground-based, satellite, and reanalysis datasets emerging an alternative and promising technique for AOD retrieval. Moreover, MLA-AOD values can be used in numerous scientific disciplines, including radiative transfer modeling, aerosol climatology, climate change, as well as air pollution applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs16071132/s1, Table S1: Information about the BSRN and AERONET stations' names, BSRN stations' acronym, location, train and test period, data availability (at 1 min resolution) and machine learning algorithm with the highest performance; Table S2: Hyperparameters that are tuned during the training of machine learning algorithms; Figure S1: Linear correlation between CAMS and AERONET total column water vapor retrievals; Figure S2: Statistical indices (MAE, RMSE and R) for Scenario 1 (a, c, e) and Scenario 2 (b, d, f) for each machine learning algorithm (MLA). The horizontal lines separate the stations based on the KG classes. The MLA with the best performance is added above each station; Figure S3: Linear correlation between ML-AOD and AERONET AOD for the stations in Equatorial (A) KG climate class; Figure S4: Linear correlation between ML-AOD and AERONET AOD for the stations in Arid (B) KG climate class; Figure S5: Linear correlation between ML-AOD and AERONET AOD for the stations in Warm temperate (C) KG climate class; Figure S6: Linear correlation between ML-AOD and AERONET AOD for the stations in Snow (D) KG climate class; Figure S7: Linear correlation between ML-AOD and AERONET AOD for the stations in Polar (E) KG climate class; Figure S8: AERONET, ML-AOD, and MODIS AOD550nm daily variability for specific months at (a) TAM, (b) XIA, (c) CAR, (d) TOR, (e) DAR, and (f) BOS sites. The second y-axis shows the daily variability of $a_{440-870nm}$.

Appendix A

This appendix presents the MLAs in more detail, aiming to describe the main ideas behind the algorithms.

- **Light Gradient Boosting Machine (LGBM)**

Gradient Boosting Machines (GBM) [76] constitute an ensemble technique that applies iteratively multiple weak learners (decision trees) in an additive and sequential form. GBM can be used either for regression or classification. More specifically, GBM is based on boosting techniques where the basic idea is to combine weak learners to build a strong learner that can efficiently predict unknown (testing) data. In this study, a modified GBM algorithm is used: the LGBM [77]. The main difference between LGBM and traditional GBM algorithm is how the decision trees are processed. LGBM uses leaf-wise tree growth instead of depth-wise tree growth (applied by GBM), leading to faster training, lower use of memory, better accuracy, and easy handling of large datasets.

- **Random Forest (RF)**

The RF algorithm forms an ensemble technique, combining the outputs from multiple decision trees [78]. RF differs compared to GBM, since the individual trees are handled using a bagging method. The predictions are configured for regression by the weighted ensemble mean of all decision tree outputs. Multiple decision trees are used as base learning models, created with bootstrapping, including a random set of training data and features for each tree node.

- **Multivariate Adaptive Regression Splines (MARS)**

MARS is a non-parametric statistical technique that can model either linear or nonlinear-related variables onto a multidimensional space to estimate a target output [79]. During training, the data are divided into piecewise linear segments (splines) of differing gradi-

ents [80]. MARS forms a linear combination of basis functions (linear, Hinge, etc.). More specifically, MARS includes two critical stages. First, a stepwise search of the possible terms that minimizes an error loss function is performed in the forward pass. The former stage often leads to quite a complex and overfitted model. Then, the backward phase follows to improve the model performance by pruning the less significant terms. The pruning process is based on minimizing the generalized cross-validation (GCV) score to select the optimal model configurations.

- **K-Nearest Neighbors (KNN)**

    KNN constitutes an algorithm that is widely used in classification tasks. Nevertheless, KNN can also be used efficiently for regression. The core concept behind the KNN algorithm is to group the training data into coherent clusters (neighbors) that can be easily identified from the testing dataset. The Euclidean distance between the point and other already known classified points is calculated to determine a given point's nearest neighbor. The selection of the optimal number of nearest neighbors controls the models' complexity and performance, making the KNN one of the simplest MLAs.

- **Artificial Neural Network (ANN)**

    ANN is generally a sequence of interacting layers consisting of neurons (nodes). There are several types of ANNs currently used in the ML community. A fully connected neural network [81] is selected for the presented study. The implemented ANN architecture includes one input, one output layer, and an intermediate layer consisting of two hidden layers. The neurons of the input layer are equal to the number of input features (three, here) while for the output layer the neurons correspond to the number of outputs (one, here). The number of hidden units included in the two hidden layers is based on the best model accuracy. The activation function between the input layer and the two hidden layers is the Rectified Linear Unit (ReLU), with a linear activation function between the second hidden layer and the output layer. The mean square error (MSE) is preferred as the loss function, and the AdaDelta (Adam) [82] performs the stochastic gradient-descent process for updating the model's weights at each iteration during training.

    Table A1 summarizes the functions and libraries–packages used to run MLAs.

Table A1. Packages along with their functions for the implemented machine learning algorithms.

MLA	Package	Function
LGBM	LightGBM Python	lgb
RF	scikit–learn Python	RandomForestRegressor
MARS	py-earth Python	Earth
KNN	scikit–learn Python	KNeighborsRegressor
ANN	Keras Application Programming Interface	[83]

References

1. Gomis, M.I.; Huang, M.; Leitzell, K.; Lonnoy, E.; Matthews, J.B.R.; Maycock, T.K.; Waterfield, T.; Yelekçi, O.; Yu, R.; Zhou, B. *IPCC, 2021: Climate Change 2021: The Physical Science Basis*; Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2021; *in press*. [CrossRef]
2. Palacios-Peña, L.; Jiménez-Guerrero, P.; Baró, R.; Balzarini, A.; Bianconi, R.; Curci, G.; Landi, T.C.; Pirovano, G.; Prank, M.; Riccio, A.; et al. Aerosol optical properties over Europe: An evaluation of the AQMEII Phase 3 simulations against satellite observations. *Atmos. Chem. Phys.* **2019**, *19*, 2965–2990. [CrossRef]
3. García, O.E.; Díaz, J.P.; Expósito, F.J.; Díaz, A.M.; Dubovik, O.; Derimian, Y.; Dubuisson, P.; Roger, J.-C. Shortwave radiative forcing and efficiency of key aerosol types using AERONET data. *Atmos. Chem. Phys.* **2012**, *12*, 5129–5145. [CrossRef]
4. Gkikas, A.; Obiso, V.; Pérez García-Pando, C.; Jorba, O.; Hatzianastassiou, N.; Vendrell, L.; Basart, S.; Solomos, S.; Gassó, S.; Baldasano, J.M. Direct radiative effects during intense Mediterranean desert dust outbreaks. *Atmos. Chem. Phys.* **2018**, *18*, 8757–8787. [CrossRef]
5. Korras-Carraca, M.-B.; Gkikas, A.; Matsoukas, C.; Hatzianastassiou, N. Global Clear-Sky Aerosol Speciated Direct Radiative Effects over 40 Years (1980–2019). *Atmosphere* **2021**, *12*, 1254. [CrossRef]

6. Logothetis, S.-A.; Salamalikis, V.; Kazantzidis, A. The impact of different aerosol properties and types on direct aerosol radiative forcing and efficiency using AERONET version 3. *Atmos. Res.* **2021**, *250*, 105343. [CrossRef]

7. Sengupta, M.; Xie, Y.; Lopez, A.; Habte, A.; Maclaurin, G.; Shelby, J. The National Solar Radiation Data Base (NSRDB). *Renew. Sustain. Energy Rev.* **2018**, *89*, 51–60. [CrossRef]

8. Javadnia, E.; Abkar, A.; Schubert, P. Estimation of High-Resolution Surface Shortwave Radiative Fluxes Using SARA AOD over the Southern Great Plains. *Remote Sens.* **2017**, *9*, 1146. [CrossRef]

9. Gueymard, C.A.; Habte, A.; Sengupta, M. Reducing Uncertainties in Large-Scale Solar Resource Data: The Impact of Aerosols. *IEEE J. Photovolt.* **2018**, *8*, 1732–1737. [CrossRef]

10. Kosmopoulos, P.G.; Kazadzis, S.; Taylor, M.; Raptis, P.I.; Keramitsoglou, I.; Kiranoudis, C.; Bais, A.F. Assessment of surface solar irradiance derived from real-time modelling techniques and verification with ground-based measurements. *Atmos. Meas. Tech.* **2018**, *11*, 907–924. [CrossRef]

11. Fountoukis, C.; Martín-Pomares, L.; Perez-Astudillo, D.; Bachour, D.; Gladich, I. Simulating global horizontal irradiance in the Arabian Peninsula: Sensitivity to explicit treatment of aerosols. *Sol. Energy* **2018**, *163*, 347–355. [CrossRef]

12. Shi, H.; Li, W.; Fan, X.; Zhang, J.; Hu, B.; Husi, L.; Shang, H.; Han, X.; Song, Z.; Zhang, Y.; et al. First assessment of surface solar irradiance derived from Himawari-8 across China. *Sol. Energy* **2018**, *174*, 164–170. [CrossRef]

13. Zhang, Y.; He, T.; Liang, S.; Wang, D.; Yu, Y. Estimation of all-sky instantaneous surface incident shortwave radiation from Moderate Resolution Imaging Spectroradiometer data using optimization method. *Remote Sens. Environ.* **2018**, *209*, 468–479. [CrossRef]

14. Vamvakas, I.; Salamalikis, V.; Benitez, D.; Al-Salaymeh, A.; Bouaichaoui, S.; Yassaa, N.; Guizani, A.; Kazantzidis, A. Estimation of global horizontal irradiance using satellite-derived data across Middle East-North Africa: The role of aerosol optical properties and site-adaptation methodologies. *Renew. Energy* **2020**, *157*, 312–331. [CrossRef]

15. Polo, J.; Fernández-Peruchena, C.; Salamalikis, V.; Mazorra-Aguiar, L.; Turpin, M.; Martín-Pomares, L.; Kazantzidis, A.; Blanc, P.; Remund, J. Benchmarking on improvement and site-adaptation techniques for modeled solar radiation datasets. *Sol. Energy* **2020**, *201*, 469–479. [CrossRef]

16. Salamalikis, V.; Vamvakas, I.; Blanc, P.; Kazantzidis, A. Ground-based validation of aerosol optical depth from CAMS reanalysis project: An uncertainty input on direct normal irradiance under cloud-free conditions. *Renew. Energy* **2021**, *170*, 847–857. [CrossRef]

17. Fountoulakis, I.; Kosmopoulos, P.; Papachristopoulou, K.; Raptis, I.-P.; Mamouri, R.-E.; Nisantzi, A.; Gkikas, A.; Witthuhn, J.; Bley, S.; Moustaka, A.; et al. Effects of Aerosols and Clouds on the Levels of Surface Solar Radiation and Solar Energy in Cyprus. *Remote Sens.* **2021**, *13*, 2319. [CrossRef]

18. 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.; et al. AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization. *Remote Sens. Environ.* **1998**, *66*, 1–16. [CrossRef]

19. Zender, C.S.; Miller, R.; Tegen, I. Quantifying mineral dust mass budgets: Terminology, constraints, and current estimates. *Eos Trans. Am. Geophys. Union* **2004**, *85*, 509–512. [CrossRef]

20. Textor, C.; Schulz, M.; Guibert, S.; Kinne, S.; Balkanski, Y.; Bauer, S.; Berntsen, T.; Berglen, T.; Boucher, O.; Chin, M.; et al. Analysis and quantification of the diversities of aerosol life cycles within AeroCom. *Atmos. Chem. Phys.* **2006**, *6*, 1777–1813. [CrossRef]

21. Kok, J.F.; Ridley, D.A.; Zhou, Q.; Miller, R.L.; Zhao, C.; Heald, C.L.; Ward, D.S.; Albani, S.; Haustein, K. Smaller desert dust cooling effect estimated from analysis of dust size and abundance. *Nat. Geosci.* **2017**, *10*, 274–278. [CrossRef]

22. Wei, X.; Chang, N.-B.; Bai, K.; Gao, W. Satellite remote sensing of aerosol optical depth: Advances, challenges, and perspectives. *Crit. Rev. Environ. Sci. Technol.* **2019**, *50*, 1640–1725. [CrossRef]

23. Paciorek, C.J.; Liu, Y.; Moreno-Macias, H.; Kondragunta, S. Spatio-temporal associations between GOES aerosol optical depth retrievals and ground-level PM2.5. *Environ. Sci. Technol.* **2008**, *42*, 5800–5806. [CrossRef] [PubMed]

24. Lindfors, A.V.; Kouremeti, N.; Arola, A.; Kazadzis, S.; Bais, A.F.; Laaksonen, A. Effective aerosol optical depth from pyranometer measurements of surface solar radiation (global radiation) at Thessaloniki, Greece. *Atmos. Chem. Phys.* **2013**, *13*, 3733–3741. [CrossRef]

25. Salmon, A.; Quiñones, G.; Soto, G.; Polo, J.; Gueymard, C.; Ibarra, M.; Cardemil, J.; Escobar, R.; Marzo, A. Advances in aerosol optical depth evaluation from broadband direct normal irradiance measurements. *Sol. Energy* **2021**, *221*, 206–217. [CrossRef]

26. Seo, J.; Choi, H.; Oh, Y. Potential of AOD Retrieval Using Atmospheric Emitted Radiance Interferometer (AERI). *Remote Sens.* **2022**, *14*, 407. [CrossRef]

27. Herrero-Anta, S.; Román, R.; Mateos, D.; González, R.; Antuña-Sánchez, J.C.; Herreras-Giralda, M.; Almansa, A.F.; González-Fernández, D.; Herrero del Barrio, C.; Toledano, C.; et al. Retrieval of aerosol properties from zenith sky radiance measurements. *Atmos. Meas. Tech.* **2023**, *16*, 4423–4443. [CrossRef]

28. Li, J.; Liu, R.; Liu, S.C.; Shiu, C.J.; Wang, J.; Zhang, Y. Trends in aerosol optical depth in northern China retrieved from sunshine duration data. *Geophys. Res. Lett.* **2016**, *43*, 431–439. [CrossRef]

29. Sanchez-Romero, A.; Sanchez-Lorenzo, A.; González, J.A.; Calbó, J. Reconstruction of long-term aerosol optical depth series with sunshine duration records. *Geophys. Res. Lett.* **2016**, *43*, 1296–1305. [CrossRef]

30. Wandji Nyamsi, W.; Lipponen, A.; Sanchez-Lorenzo, A.; Wild, M.; Arola, A. A hybrid method for reconstructing the historical evolution of aerosol optical depth from sunshine duration measurements. *Atmos. Meas. Tech.* **2020**, *13*, 3061–3079. [CrossRef]

31. Kazantzidis, A.; Tzoumanikas, P.; Nikitidou, E.; Salamalikis, V.; Wilbert, S.; Prahl, C. Application of Simple All-sky Imagers for the Estimation of Aerosol Optical Depth. *AIP Conf. Proc.* **2017**, *1850*, 140012. [CrossRef]

32. Román, R.; Antuña-Sánchez, J.C.; Cachorro, V.E.; Toledano, C.; Torres, B.; Mateos, D.; Fuertes, D.; López, C.; González, R.; Lapionok, T.; et al. Retrieval of aerosol properties using relative radiance measurements from an all-sky camera. *Atmos. Meas. Tech.* **2022**, *15*, 407–433. [CrossRef]

33. Scarlatti, F.; Gómez-Amo, J.L.; Valdelomar, P.C.; Estellés, V.; Utrillas, M.P. A Machine Learning Approach to Derive Aerosol Properties from All-Sky Camera Imagery. *Remote Sens.* **2023**, *15*, 1676. [CrossRef]

34. Logothetis, S.-A.; Giannaklis, C.-P.; Salamalikis, V.; Tzoumanikas, P.; Raptis, P.-I.; Amiridis, V.; Eleftheratos, K.; Kazantzidis, A. Aerosol Optical Properties and Type Retrieval via Machine Learning and an All-Sky Imager. *Atmosphere* **2023**, *14*, 1266. [CrossRef]

35. Olcese, L.E.; Palancar, G.G.; Toselli, B.M. A method to estimate missing AERONET AOD values based on artificial neural networks. *Atmos. Environ.* **2015**, *113*, 140–150. [CrossRef]

36. 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.; et al. 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.* **2016**, *16*, 8181–8191. [CrossRef]

37. Nabavi, S.O.; Haimberger, L.; Abbasi, R.; Samimi, C. Prediction of aerosol optical depth in West Asia using deterministic models and machine learning algorithms. *Aeolian Res.* **2018**, *35*, 69–84. [CrossRef]

38. Kolios, S.; Hatzianastassiou, N. Quantitative Aerosol Optical Depth Detection during Dust Outbreaks from Meteosat Imagery Using an Artificial Neural Network Model. *Remote Sens.* **2019**, *11*, 1022. [CrossRef]

39. Zbizika, R.; Pakszys, P.; Zielinski, T. Deep Neural Networks for Aerosol Optical Depth Retrieval. *Atmosphere* **2022**, *13*, 101. [CrossRef]

40. Dubovik, O.; Lapyonok, T.; Litvinov, P.; Herman, M.; Fuertes, D.; Ducos, F.; Lopatin, A.; Chaikovsky, A.; Torres, B.; Derimian, Y.; et al. GRASP: A versatile algorithm for characterizing the atmosphere. *SPIE Newsroom* **2014**, *25*, 2-1201408. [CrossRef]

41. Dubovik, O.; King, M.D. A flexible inversion algorithm for retrieval of aerosol optical properties from sun and sky radiance measurements. *J. Geophys. Res.* **2000**, *105*, 20673–20696. [CrossRef]

42. 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.; et al. 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.* **2019**, *12*, 169–209. [CrossRef]

43. Ohmura, A.; Gilgen, H.; Hegner, H.; Müller, G.; Wild, M.; Dutton, E.G.; Forgan, B.; Fröhlich, C.; Philipona, R.; Heimo, A.; et al. Baseline Surface Radiation Network (BSRN/WCRP): New Precision Radiometry for Climate Research. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 2115–2136. [CrossRef]

44. Driemel, A.; Augustine, J.; Behrens, K.; Colle, S.; Cox, C.; Cuevas-Agulló, E.; Denn, F.M.; Duprat, T.; Fukuda, M.; Grobe, H.; et al. Baseline Surface Radiation Network (BSRN): Structure and data description (1992–2017). *Earth Syst. Sci. Data* **2018**, *10*, 1491–1501. [CrossRef]

45. Yang, D.; Yagli, G.M.; Quan, H. Quality Control for Solar Irradiance Data. In Proceedings of the 2018 IEEE Innovative Smart Grid Technologies—Asia (ISGT Asia), Singapore, 22–25 May 2018; pp. 208–213. [CrossRef]

46. Yang, D. SolarData package update v1.1: R functions for easy access of Baseline Surface Radiation Network (BSRN). *Sol. Energy* **2019**, *188*, 970–975. [CrossRef]

47. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* **2006**, *15*, 259–263. [CrossRef] [PubMed]

48. Bryant, C.; Wheeler, N.R.; Rubel, F.; French, R.H. Kgc: Koeppen-Geiger Climatic Zones. R Package Version 1.0.0.2. 2017. Available online: https://cran.r-project.org/package=kgc (accessed on 22 March 2024).

49. Bright, J.M.; Gueymard, C.A. Climate-specific and global validation of MODIS Aqua and Terra aerosol optical depth at 452 AERONET stations. *Sol. Energy* **2019**, *183*, 594–605. [CrossRef]

50. Gueymard, C.A.; Yang, D. Worldwide validation of CAMS and MERRA-2 reanalysis aerosol optical depth products using 15 years of AERONET observations. *Atmos. Environ.* **2020**, *225*, 117216. [CrossRef]

51. Wei, J.; Li, Z.; Peng, Y.; Sun, L. MODIS Collection 6.1 aerosol optical depth products over land and ocean: Validation and comparison. *Atmos. Environ.* **2019**, *201*, 428–440. [CrossRef]

52. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.-M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys.* **2019**, *19*, 3515–3556. [CrossRef]

53. Popp, T.; de Leeuw, G.; Bingen, C.; Brühl, C.; Capelle, V.; Chedin, A.; Clarisse, L.; Dubovik, O.; Grainger, R.; Griesfeller, J.; et al. Development, production and evaluation of aerosol climate data records from european satellite observations (Aerosol_cci). *Remote Sens.* **2016**, *8*, 421. [CrossRef]

54. Buchard, V.; Randles, C.A.; da Silva, A.M.; Darmenov, A.; Colarco, P.R.; Govindaraju, R.; Ferrare, R.; Hair, J.; Beyersdorf, A.J.; Ziemba, L.D.; et al. The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies. *J. Clim.* **2017**, *30*, 6851–6872. [CrossRef]

55. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef] [PubMed]
56. Randles, C.A.; da Silva, A.M.; Buchard, V.; Colarco, P.R.; Darmenov, A.; Govindaraju, R.; Smirnov, A.; Holben, B.; Ferrare, R.; Hair, J.; et al. The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. *J. Clim.* **2017**, *30*, 6823–6850. [CrossRef] [PubMed]
57. Levy, R.C.; Mattoo, S.; Munchak, L.A.; Remer, L.A.; Sayer, A.M.; Patadia, F.; Hsu, N.C. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* **2013**, *6*, 2989–3034. [CrossRef]
58. Levy, R.C.; Remer, L.A.; Dubovik, O. Global aerosol optical properties and application to Moderate Resolution Imaging Spectroradiometer aerosol retrieval over land. *J. Geophys. Res. Atmos.* **2007**, *112*, D13210. [CrossRef]
59. Levy, R.C.; Remer, L.A.; Mattoo, S.; Vermote, E.F.; Kaufman, Y.J. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. *J. Geophys. Res. Atmos.* **2007**, *112*, D13211. [CrossRef]
60. Levy, R.C.; Remer, L.A.; Kleidman, R.G.; Mattoo, S.; Ichoku, C.; Kahn, R.; Eck, T.F. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. *Atmos. Chem. Phys.* **2010**, *10*, 10399–10420. [CrossRef]
61. Remer, L.A.; Tanre, D.; Kaufman, Y.J.; Ichoku, C.; Mattoo, S.; Levy, R.; Chu, D.A.; Holben, B.; Dubovik, O.; Smirnov, A.; et al. Validation of MODIS aerosol retrieval over ocean. *Geophys. Res. Lett.* **2002**, *29*, MOD3.1–MOD3.4. [CrossRef]
62. Remer, L.A.; Kaufman, Y.J.; Tanre, D.; Mattoo, S.; Chu, D.A.; Martins, J.V.; Li, R.R.; Ichoku, C.; Levy, R.C.; Kleidman, R.G.; et al. The MODIS aerosol algorithm, products, and validation. *J. Atmos. Sci.* **2005**, *62*, 947–973. [CrossRef]
63. Remer, L.A.; Kleidman, R.G.; Levy, R.C.; Kaufman, Y.J.; Tanré, D.; Mattoo, S.; Martins, J.V.; Ichoku, C.; Koren, I.; Yu, H.; et al. Global aerosol climatology from the MODIS satellite sensors. *J. Geophys. Res. Atmos.* **2008**, *113*, D14S07. [CrossRef]
64. Sayer, A.M.; Hsu, N.C.; Bettenhausen, C.; Jeong, M.-J. Validation and uncertainty estimates for MODIS Collection 6 "Deep Blue" aerosol data. *J. Geophys. Res.* **2013**, *118*, 7864–7873. [CrossRef]
65. Hsu, N.C.; Tsay, S.-C.; King, M.D.; Herman, J.R. Aerosol Properties Over Bright-Reflecting Source Regions. *IEEE Trans. Geosci. Remote* **2004**, *42*, 557–569. [CrossRef]
66. Reno, M.J.; Hansen, C.W. Identification of periods of clear sky irradiance in time series of GHI measurements. *Renew. Energy* **2016**, *90*, 520–531. [CrossRef]
67. Kasten, F.; Young, A.T. Revised optical air mass tables and approximation formula. *Appl. Opt.* **1989**, *28*, 4735–4738. [CrossRef] [PubMed]
68. Kaufman, Y.J.; Tanre, D.; Remer, L.; Vermote, E.; Chu, A.; Holben, B.N. Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. *J. Geophys. Res. Atmos.* **1997**, *102*, 17051–17067. [CrossRef]
69. Mayer, B.; Kylling, A. Technical note: The libRadtran software package for radiative transfer calculations—Description and examples of use. *Atmos. Chem. Phys.* **2005**, *5*, 1855–1877. [CrossRef]
70. Emde, C.; Buras-Schnell, R.; Kylling, A.; Mayer, B.; Gasteiger, J.; Hamann, U.; Kylling, J.; Richter, B.; Pause, C.; Dowling, T.; et al. The libRadtran software package for radiative transfer calculations (version 2.0.1). *Geosci. Model Dev.* **2016**, *9*, 1647–1672. [CrossRef]
71. Ruiz-Arias, J.A.; Gueymard, C.A. Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface. *Sol. Energy* **2018**, *168*, 10–29. [CrossRef]
72. Ran, L.; Deng, Z.-Z.; Wang, P.C.; Xia, X.-A. Black carbon and wavelength-dependent aerosol absorption in the North China Plain based on two-year aethalometer measurements. *Atmos. Environ.* **2016**, *142*, 132–144. [CrossRef]
73. Gueymard, C.A. Direct solar transmittance and irradiance predictions with broadband models. Part I: Detailed theoretical performance assessment. *Sol. Energy* **2003**, *74*, 355–379. [CrossRef]
74. Guirado, C.; Cuevas, E.; Cachorro, V.E.; Toledano, C.; Alonso-Pérez, S.; Bustos, J.J.; Basart, S.; Romero, P.M.; Camino, C.; Mimouni, M.; et al. Aerosol characterization at the Saharan AERONET site Tamanrasset. *Atmos. Chem. Phys.* **2014**, *14*, 11753–11773. [CrossRef]
75. Logothetis, S.-A.; Salamalikis, V.; Kazantzidis, A. Aerosol classification in Europe, Middle East, North Africa and Arabian Peninsula based on AERONET Version 3. *Atmos. Res.* **2020**, *239*, 104893. [CrossRef]
76. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
77. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 9. *Adv. Neural Inf. Process. Syst.* **2017**, 3147–3155. [CrossRef]
78. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
79. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
80. Zhang, W.; Goh, A.T.C. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geosci. Front.* **2016**, *7*, 45–52. [CrossRef]
81. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef]

82.   Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. Available online: http://arxiv.org/abs/1412.6980 (accessed on 29 January 2017).

83.   Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 22 March 2024).