



Bias correction of MODIS retrieved Aerosol Optical Depth using Machine Learning and Deep Learning Techniques

M. Anitha & Lakshmi Sutha Kumar

To cite this article: M. Anitha & Lakshmi Sutha Kumar (2025) Bias correction of MODIS retrieved Aerosol Optical Depth using Machine Learning and Deep Learning Techniques, International Journal of Remote Sensing, 46:22, 8675-8710, DOI: [10.1080/01431161.2025.2571235](https://doi.org/10.1080/01431161.2025.2571235)

To link to this article: <https://doi.org/10.1080/01431161.2025.2571235>



View supplementary material [↗](#)



Published online: 28 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 58



View related articles [↗](#)



View Crossmark data [↗](#)



Bias correction of MODIS retrieved Aerosol Optical Depth using Machine Learning and Deep Learning Techniques

M. Anitha  and Lakshmi Sutha Kumar

Department of ECE, NIT Puducherry, Karaikal, India

ABSTRACT

Aerosols influence climate by interacting with solar radiation and altering cloud properties. Satellite-based Aerosol Optical Depth (AOD) measurements, such as MODIS (Moderate Resolution Imaging Spectroradiometer), provide wide coverage but suffer from biases due to atmospheric variability and algorithm limitations. This study aims to enhance MODIS AOD retrievals using AERONET (AErosol RObotic NETwork) AOD as a reference, focusing on the Kanpur station. Essential meteorological variables from ECMWF (European Centre for Medium-Range Weather Forecasts) reanalysis were used to correct AOD biases from MODIS Collection 6.1 (Terra: 2001–2022, Aqua: 2002–2022), covering Dark Target ($DT_{3\text{ km}}$ and $DT_{10\text{ km}}$) and Deep Blue ($DB_{10\text{ km}}$) products. Machine Learning algorithms, including AdaBoost (AdaB), Decision Tree (DTree), Random Forest (RF), Support Vector Regression (SVR), XGBoost (XGB), LightGBM (LGBM), CatBoost (CatB), and a Deep Learning model, namely Artificial Neural Networks (ANN), were employed. Advanced techniques such as 10-fold cross-validation and GridSearchCV were used for hyperparameter tuning and model stability. Performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), correlation (R), Mean Absolute Percentage Error (MAPE), Mean Bias (MB), and percentage of data within the expected error envelope (EE). Among the tested models, CatB consistently outperformed others, achieving the best balance of accuracy and reliability for both Terra and Aqua aerosol products. The CatB model improved the correlation values for Aqua $DT_{3\text{ km}}$, Aqua $DT_{10\text{ km}}$, Aqua $DB_{10\text{ km}}$, Terra $DT_{3\text{ km}}$, Terra $DT_{10\text{ km}}$, and Terra $DB_{10\text{ km}}$ by 9.42%, 18.92%, 17.55%, 6.21%, 9.13%, and 10.69% respectively. Similarly, the percentage of data within EE also improved in the ranges between 19.75% (Terra $DB_{10\text{ km}}$) and 27.34% (Aqua $DT_{10\text{ km}}$).

ARTICLE HISTORY

Received 29 May 2025

Accepted 01 October 2025


KEYWORDS

Aerosol Optical Depth; Dark Target; Deep Blue; Machine Learning; Deep Learning

1. Introduction

Air pollution is a serious environmental risk, with more than 80% of urban populations suffering from air quality levels exceeding WHO (World Health Organization) guidelines. It causes serious health issues such as heart disease, lung injury, and respiratory diseases. India is confronting severe air quality deterioration, significantly affecting premature

CONTACT M. Anitha  anithasekaran9@gmail.com  Department of Electronics and Communication Engineering, National Institute of Technology Puducherry, Karaikal, Puducherry, India

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/01431161.2025.2571235>.

© 2025 Informa UK Limited, trading as Taylor & Francis Group

death rates (Anitha and Kumar 2023a). Aerosols are atmospheric suspensions of solid or liquid particles that typically have aerodynamic sizes less than $100\text{ }\mu\text{m}$ (Anitha and Kumar 2023b; Cuneo, Ulke, and Cerne 2022). It originates from man-made (like burning and industrial pollutants) and natural (like dust storms and sea spray) sources. They affect the climate of the Earth by scattering and absorbing radiation, altering cloud characteristics, and modifying precipitation patterns (Choi, Lee, and Park 2021; Fan et al. 2021; Liu et al. 2024; Olcese, Palancar, and Toselli 2014; Sangura et al. 2025). They pose a significant challenge to climate research today through their effect on the Earth's radiation balance, air quality, and human health (Gong et al. 2014; Kim et al. 2021; Kumar et al. 2022). Aerosol characteristics have been observed with satellite, airborne, and ground-based measurements during the last two decades (Sabetghadam et al. 2021). Their short atmospheric lifetimes and spatial-temporal fluctuations complicate global characterization (Giles et al. 2012). Due to their dangerous health effects, continuous monitoring and control of aerosols are essential for public health and environmental protection.

Aerosol Optical Depth (AOD) measures how much aerosols reduce light transmission via absorption or scattering. A value of 0.01 suggests a very clean atmosphere, and a value of 0.4 indicates a hazy atmospheric condition (Global Monitoring Laboratory 2019). It is essential in various applications, including monitoring aerosol sources, volcanic eruptions, biomass burning, and radiative transfer model computations, correlating well with $PM_{2.5}$ (particles having a radius less than $2.5\text{ }\mu\text{m}$) levels (Weber et al. 2010). Researchers have undertaken significant efforts to measure AOD, which they can obtain from ground-based and satellite-based instruments. Surface measurements, such as AErosol RObotic NETwork (AERONET) (Dubovik et al. 2000), provide precise and frequent aerosol data but are limited to specific locations (Gao et al. 2016). Satellite sensors such as Moderate Resolution Imaging Spectroradiometer (MODIS), Multi-angle Imaging Spectro-Radiometer (MISR), Sea-viewing Wide Field-of-view Sensor (SeaWiFS), Ozone Monitoring Instrument (OMI), and others (Albayrak et al. 2013) that provide global aerosol measurements but are based on assumptions of surface reflectance and aerosol characteristics, hence impacting accuracy.

Though satellites are inexpensive for aerosol monitoring, AERONET measurements are more accurate. Satellite-based AOD observations, especially from MODIS on NASA's Terra and Aqua satellites, provide global coverage and complement AERONET's accurate ground-based measurements. Researchers are making significant efforts to compare and collocate these diverse datasets, but biases between them still exist. The possible reasons for uncertainty in MODIS AOD retrievals are instrumental errors, atmospheric variability, and algorithmic assumptions. Calculating aerosol radiative forcing is challenging due to uncertainties from varying conditions and measurement errors. Despite these limitations, satellite observations enhance our understanding of aerosol variations and support large-scale monitoring. Several global studies comparing MODIS AOD with AERONET measurements have found strong correlations across regions (More et al. 2013; Remer et al. 2008).

Some previous studies that validated MODIS aerosol products in India using correlation (R) are as follows. Misra, Jayaraman, and Ganguly (2008, 2015) validated MODIS aerosol products over Ahmedabad using Microtops sun photometer data. Misra, Jayaraman, and Ganguly (2008) found $R = 0.71\text{--}0.75$, while Misra, Jayaraman, and Ganguly (2015) reported Aqua DT (Dark Target) ($R = 0.69$), Terra DT ($R = 0.55$), Aqua DB (Deep Blue) ($R = 0.50$), and Terra DB ($R = 0.43$). Prasad and Singh (2007) assessed MISR and MODIS Terra AOD using

AERONET data from 2001–2005 over Kanpur ($R^2 = 0.7$), while Tripathi et al. (2005) found $R^2 = 0.71$ – 0.72 for MODIS vs. AERONET AOD in 2004. More et al. (2013) compared AERONET, Microtops, and MODIS AOD over Pune, reporting $R = 0.62$ – 0.93 and WEE (Percentage of data within the expected error boundary) = 68%–84%. Vijayakumar et al. (2018) analysed AOD and PWV (Precipitable Water Vapour) over Pune (2005–2015), using MODIS, ECMWF (European Centre for Medium-Range Weather Forecasts) reanalysis, and AERONET data, finding $R = 0.73$ – 0.79 . In the Delhi National Capital Region (NCR), Sharma et al. (2021) evaluated the C6.1 Terra and Aqua MODIS DT AOD at 10 km ($DT_{10\text{ km}}$) and 3 km ($DT_{3\text{ km}}$) resolutions, along with DB AOD at 10 km ($DB_{10\text{ km}}$), using AERONET observations over 10 years (2010–2019). They reporting for Terra: $DB_{10\text{ km}}$ ($R = 0.44$), $DT_{3\text{ km}}$ ($R = 0.74$), $DT_{10\text{ km}}$ ($R = 0.77$), and for Aqua: $DB_{10\text{ km}}$ ($R = 0.7$), $DT_{3\text{ km}}$ ($R = 0.79$), $DT_{10\text{ km}}$ ($R = 0.72$). Mhawish et al. (2017) validated MODIS C6 AOD algorithms ($DB_{10\text{ km}}$, $DT_{3\text{ km}}$, $DT_{10\text{ km}}$, and combined DT-DB AOD at 10 km ($DTB_{10\text{ km}}$) using AERONET data from six stations over the Indo-Gangetic Plain (IGP) (2006–2015), with $R = 0.7$ – 0.8 and WEE = 51.37%–61.29%. Mangla, Indu, and Chakra (2020) compared AOD from OMI, MISR, and MODIS with ground-based data (2010–2017), finding $R^2 = 0.7$. Although the MODIS and AERONET AOD have a stronger correlation in these studies, AOD can yet be improved (Lanzaco et al. 2016).

Researchers mitigate AOD biases by cross-calibrating radiances and applying aerosol retrieval algorithms to multisensory data (Albayrak et al. 2013). Zhang and Reid (2006) developed empirical corrections for oceanic AOD using AERONET. Scientists periodically adjust the surface reflectance models, AOD retrieval algorithms, and terrains to improve AOD accuracy over land (Lanzaco et al. 2016, 2017). The nonlinear nature of AOD proves to be challenging for traditional linear techniques. Therefore, researchers turn to Machine Learning (ML) and Deep Learning (DL) for better MODIS-AERONET consistency. Such data-driven approaches do not have prior assumptions and perform well in varied retrieval scenarios. ML algorithms outperform physical models such as DT and DB when sufficient training data is available, capturing complex patterns for accurate AOD correction (M. Wang et al. 2023). ML models can also estimate AOD over regions without ground-based retrievals and are becoming increasingly popular in aerosol science (Hirtl et al. 2014). The following are several previous works that adjusted the AOD bias using various ML and DL algorithms.

To increase the precision of AOD retrieval from MODIS data using the DT technique, Hang et al. (2018) suggest two hybrid frameworks (serial and parallel) based on Ridge Regression (RR), along with the use of calibrated radiance product, geolocation product, aerosol product, and cloud mask product. Experiments on 3093 collocated observations from 10 stations in China demonstrate that both hybrid approaches perform better in retrieval than the RR model and the standalone DT algorithm. For instance, when MODIS Aerosol Optical Thickness (AOT) data is adjusted using a NN (Neural Network), biases and mistakes are compensated for, enhancing correlations with AERONET data by 4–6% and increasing the proportion of accurate data points within the predicted error envelope by roughly 10% (Albayrak et al. 2013). In addition to using ML to estimate global $PM_{2.5}$ levels, Malakar et al. (2012) use an NN to predict AERONET AOD from MODIS data and investigate reasons for bias. A post-processing correction model based on Random Forest (RF) and NN was created by Lipponen et al. (2021) to increase accuracy by addressing biases through the use of auxiliary data from 2014 to 2018. Using these models, the percentage of MODIS AOD samples within the anticipated error envelope increased from 63% to 85%,

providing a computationally efficient approach to enhancing existing satellite aerosol datasets. To examine the persistent bias between AERONET and MODIS AOD, Support Vector Machines (SVM) and NN were employed by Lary et al. (2009). The results imply that MODIS AOD biases could be related to surface type, surface reflectance, or the covariance between aerosol properties and surface features. Just et al. (2018) compare three ML methods (RF, Gradient Boosting (GB), and XGBoost (XGB)) to correct measurement errors in the MAIAC (Multi-Angle Implementation of Atmospheric Correction) AOD product from Aqua and Terra satellites, using data from 79 AERONET stations across the North-eastern /Mid-Atlantic U.S.A.. In this, XGB outperformed the other methods, reducing the Root Mean Squared Error (RMSE) by 43% and 44% for Aqua and Terra and improving the correlation between AOD and daily $PM_{2.5}$ monitors by up to 10%.

Using MODIS satellite data, Lemmouchi et al. (2023) offer four supervised ML regression methods, such as multiple linear regression, RF, XGB and Artificial Neural Networks (ANN), to enhance CHIMERE-simulated AOD over North Africa and the Arabian Peninsula using satellite and geophysical data. All models showed fewer biases and errors, but RF displayed fewer spatial artefacts. The technique improves daily AOD accuracy, especially in high-aerosol regions like the Sahara. The above works use the MODIS retrieved data as features and AERONET AOD for the AOD bias correction. Similarly, the following papers use MODIS AOD and meteorological variables, with AERONET AOD as the target variable, to reduce satellite AOD bias: To bring MODIS AOD values closer to AERONET measurements, Lanzaco et al. (2016) proposed two ML techniques, such as ANN and SVR (Support Vector Regression). The method successfully reduced biases and outliers by increasing the percentage of data within the MODIS predicted error from 57% to 91% when researchers applied it to nine South American stations, and both algorithms performed well. Lanzaco et al. (2017) improved the accuracy of over 62% of South America by using ANN and SVM to adjust MODIS AOD readings in areas remote from AERONET bases. For Terra and Aqua, the percentage of data within the MODIS error rose to 38% and 86%, respectively. Using MODIS AOD, meteorological reanalysis, and AERONET data, M. Wang et al. (2023) use SVR and ANN to rectify overestimated MODIS AOD values in the Beijing region. With R^2 , RMSE, and slope values of 0.88, 0.12, and 0.97, respectively, across 20 years of data (2001–2019), the ANN model fared better than SVR.

Despite significant advancements in MODIS AOD bias correction using ML, several research gaps remain. The literature survey indicates that researchers can correct MODIS AOD bias using satellite-retrieved data or meteorological parameters. However, most of the existing ML models only take a part of the spectral values from the MODIS satellite retrieval as data, resulting in the loss of adequate information. Moreover, dataset preparation is still a big challenge because of the need for multiple satellite products and complicated collocation processes. In some work, they used reflectance products, and the upscaling or downscaling became tedious. On the other hand, using meteorological variables as features in MODIS AOD bias correction ensures effortless dataset preparation. Many approaches also filter outliers, potentially disregarding crucial real-world AOD variability. Most existing studies primarily focus on correcting $DT_{10\text{ km}}$ AOD, with limited attention to $DB_{10\text{ km}}$ and $DT_{3\text{ km}}$ retrievals, leading to incomplete bias correction across MODIS products.

Furthermore, a majority of research has only utilized a subset of ML algorithms and ignored recently developed, more effective ones, such as AdaBoost (AdaB),

LightGBM (LGBM), and CatBoost (CatB), which might provide better results. Few papers have used feature selection techniques, and the significance of different input variables has not been rigorously evaluated on a range of ML models. Additionally, strict model validation procedures like 10-fold cross-validation in conjunction with GridSearchCV for the best hyperparameter tuning have not been done extensively to warrant generalizability. Previous methods generally don't incorporate rigorous performance assessment across multiple measures like RMSE, Mean Absolute Error (MAE), R, Mean Absolute Percentage Error (MAPE), WEE and Mean Bias (MB), which are crucial for a comprehensive understanding of model performance. Earlier research has predominantly used small datasets with limited long-term evaluations that effectively capture seasonal and interannual variations. Addressing these gaps, this study proposes a novel comprehensive ML/DL-based bias-correction framework for MODIS AOD, incorporating advanced ML models, extensive feature selection, robust validation, and multiple evaluation metrics to enhance MODIS AOD retrieval accuracy, particularly over Kanpur, but with scope for extension to other areas. The significant contribution of this work is as follows:

- (1) Incorporating key meteorological variables from ECMWF, along with the year and DOY (Day of the Year), the model effectively captures temporal and atmospheric influences over the entire MODIS measurement period (Terra: 2001–2022, Aqua: 2002–2022). Using the most recent MODIS Collection 6.1 (C6.1), which fixes earlier calibration issues, guarantees more accurate aerosol retrievals. In contrast to conventional techniques, our work preserves outliers during ML/DL training, improving model resilience to actual AOD variability.
- (2) This work employs advanced models like AdaB, Decision Tree (DTree), RF, SVR, XGB, LGBM, CatB and ANN to provide a scalable and automated way to correct local biases and outliers present in the $DT_{3\text{ km}}$, $DT_{10\text{ km}}$ and $DB_{10\text{ km}}$ AOD datasets of both Aqua and Terra satellites. The framework improves MODIS AOD retrieval reliability, which makes it more appropriate for climate and air quality applications.
- (3) To ensure robust insights for AOD prediction, an extensive feature selection method utilizing several ML techniques is employed. Selecting features appropriate for various prediction models is made possible by combining traditional importance, permutation importance, and SHAP (SHapley Additive exPlanations) values across models to provide a thorough knowledge of feature contributions.
- (4) To improve training stability, prevent overfitting, and enhance model generalization, the 10-fold cross-validation is used along with GridSearchCV, which selects the best hyperparameters for each ML model. In addition, some more regularization techniques are employed, such as batch normalization, L2 regularization, and dropout, to improve the ANN model's performance.
- (5) To provide a comprehensive comparison framework for future AOD bias-correction studies, models are rigorously evaluated using RMSE, MAE, R, MAPE, MB, WEE, AEE (percentage of data above the error envelope boundary), and BEE (percentage of data below the error envelope boundary).
- (6) The research offers one of the first long-term ML-based AOD bias correction analyses for Kanpur, potentially applicable to other regions.

The remaining Sections of this paper are organized as follows: [Section 2](#) describes the study area and dataset used. [Section 3](#) explains the details of the methodology, ML and DL algorithms, and performance metrics. [Section 4](#) presents the results and discussions, and finally, the work is concluded in [Section 5](#).

2. Study site and datasets used

This Section describes the study site and various datasets from the AERONET, MODIS, and ECMWF reanalysis models.

2.1. Study location

An industrial city in central India, Kanpur is exposed to high aerosol loads and increasing pollution because of accelerated industrialization and urbanization. The Yamuna and Ganga rivers bound the city. Frequent dust storms in the IGP aggravate the pollution, with emissions from factories, power plants, vehicles, and biomass burning as significant contributors. Biomass burning is due to agricultural residues, primarily affecting aerosol composition in this region (Annapurna, Anitha, and Kumar 2024). The analysis of this site provides valuable data for aerosol source identification and control. Kanpur AERONET station (26.51278° N, 80.23164° E, Elevation: 123 m), 17 km away from the city centre, is chosen for this study because of its high AOD values and the presence of two decades of data. [Figure 1](#) shows the Kanpur AERONET station in the Kanpur Nagar district, Uttar Pradesh, India.

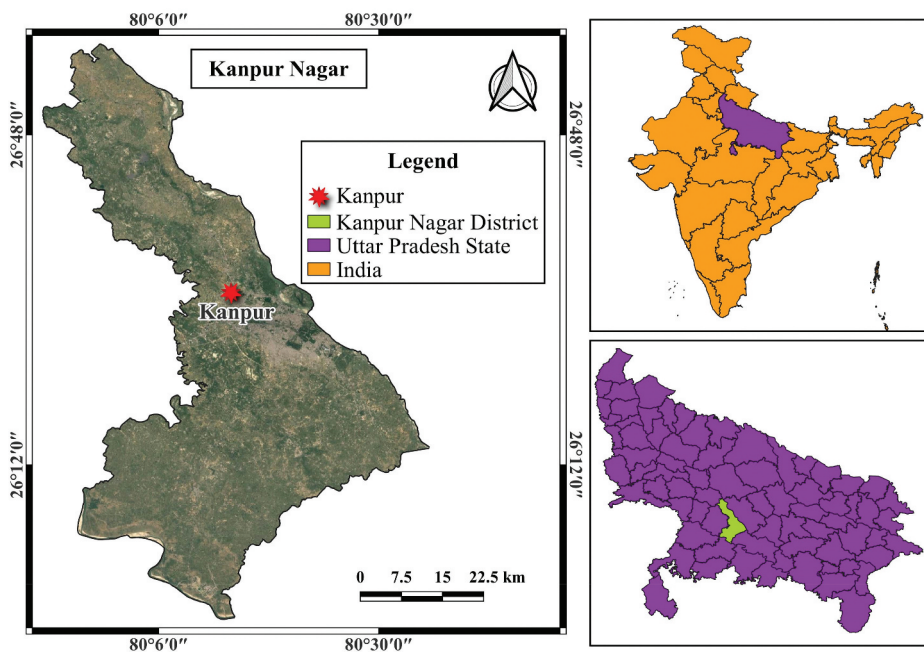


Figure 1. Map of the region of study.

2.2. Dataset overview and description

2.2.1. AERONET data

NASA's (National Aeronautics and Space Administration) AERONET network employs a sun photometer to observe aerosols over various wavelengths. The French CE-318 automatic solar scattered/direct radiometer measures scattered and direct radiation (Anitha and Kumar 2020; Fan et al. 2021). It provides broad spatial and temporal coverage for aerosol monitoring by point measurements. India has 24 AERONET sites among 400 installations worldwide in 50 nations (Mohan, Manisekaran, and Kumar 2021). AERONET determines AOD using the measured sun extinction values, removing attenuation produced by gaseous contaminants, ozone absorption, and Rayleigh scattering by applying the Lambert-Beer-Bouguer law. The computed error range for shorter wavelengths is ± 0.02 , while for longer wavelengths (above 440 nm), the margin is ± 0.01 (Tan et al. 2015). Aerosol parameters were measured by the sun photometer, which is necessary to calculate aerosol properties like Single Scattering Albedo (SSA), Angstrom Exponent (AE), Volume Size Distribution (VSD), and columnar phase function. Maritime Aerosol Network (MAN) observes aerosols in oceanic and aquatic areas. AERONET data are employed to validate satellite-based aerosol measurements that are useful in global atmospheric properties. The AERONET website (https://aeronet.gsfc.nasa.gov/new_web/webtool_aod_v3.html) provides access to the monthly and daily averaged AOD data. The AERONET offers three different levels of data: Level 1, Level 1.5, and Level 2 (Yusuf et al. 2021). This paper uses version 3 level 2 aerosol data that was cloud-screened, high-quality, and guaranteed for 2001 to 2022. MODIS gives AOD data at 550 nm; however, the adjacent available wavelength for AERONET AOD is 500 nm. Thus, for comparison and validation, this AOD needs to be interpolated to 550 nm using Equations (1) and (2) (Abd Jalal, Asmat, and Ahmad 2015).

$$(\text{AOD})_{550\text{nm}} = (\text{AOD})_{500\text{nm}} \times \left(\frac{550}{500} \right)^{-\alpha} \quad (1)$$

$$\alpha = - \frac{\ln\left(\frac{\tau_1}{\tau_2}\right)}{\ln\left(\frac{\lambda_1}{\lambda_2}\right)} \quad (2)$$

where α denotes the AE in the 440–675 nm wavelength range. τ_1 and τ_2 represents AOD observations at wavelengths 675 nm (λ_1) and 440 nm (λ_2) respectively.

2.2.2. MODIS data

Satellite remote sensing offers widespread spatial coverage. MODIS, an image sensor, was launched by NASA on the Terra satellite in 1999 and the Aqua satellite in 2002 (Li et al. 2019). It records data in 36 spectral bands with wavelengths from 0.4 μm to 14.4 μm and has various spatial resolutions: 29 bands at 1 km, two at 250 m, and five at 500 m. It scans the entire planet within one to two days with a swath width of 2330 km. Aqua crosses the equator at approximately 1:30 pm local time and Terra at 10:30 am, concentrating on observing large-scale Earth and climate phenomena (Zaman et al. 2021). MODIS retrieval algorithms play a significant role in global aerosol monitoring, particularly in desert and urban regions. To improve the accuracy of AOD data derived from MODIS, the retrieval

methods on both Aqua and Terra satellites have undergone regular updates. The updates also involve the most recent data on aerosol models, surface reflectance, and global cloud-masking processes.

These methods, including DT and DB, use distinct algorithms for AOD retrieval based on surface characteristics. DT is designed for dark vegetated land and ocean surfaces, relying on precomputed lookup tables (LUT) for aerosol and surface properties. It retrieves AOD at 10 km and 3 km spatial resolutions, assuming low surface albedo (Akoshile et al. 2019). The retrieval algorithm groups 20×20 pixels with a 500 m resolution at 0.47, 0.65, and 2.13 μm channels in a $10 \text{ km} \times 10 \text{ km}$ retrieval box. The remaining 20% darkest and 50% brightest pixels are eliminated, along with all pixels that reflect cloud, desert, snow/ice, or inland water. The algorithm uses a dynamic relationship between visible (0.47 and 0.65 μm) and infrared (2.13 μm) channels to parameterize the surface reflectance of the two visible channels. The algorithm applies aerosol models based on geolocation and season, matching radiance values in the LUT to retrieve spectral AOD. Specifically for land areas, the DT technique uses one coarse aerosol model and three different fine aerosol models (low, moderate, and highly absorbing). The algorithm arranges 36 pixels in a $3 \text{ km} \times 3 \text{ km}$ retrieval box for $DT_{3 \text{ km}}$, in contrast to $DT_{10 \text{ km}}$ (Anitha and Kumar 2024). AOD is produced by averaging a maximum of 11 pixels for a 3 km distance and 120 pixels for a 10 km distance. DT has an error range of $\pm (0.05 + 15\% \text{ AOD}_{\text{AERONET}})$ for the 10 km product, and it is $\pm (0.05 + 20\% \text{ AOD}_{\text{AERONET}})$ for the 3 km product (Sharma et al. 2021).

The DB algorithm retrieves AOD at 10 km resolution, targeting high-reflectance areas in bright land surfaces such as deserts, semiarid and urban areas. It retrieves aerosol parameters utilizing deep-blue wavelengths (0.412, 0.47 or 0.65 μm). The clear-sky pixels are retrieved by this method at a 1 km resolution, and the results are then aggregated to 10 km resolution. According to location, season, and land cover category, surface reflectance is distributed to the clear-sky pixels based on the LUT in visible bands at 10 km resolution (Shi et al. 2018). DB estimates AOD over bright and vegetation surfaces using the Normalized Difference Vegetation Index (NDVI) and reflectance ratios with an error envelope of $\pm (0.03 + 20\% \text{ AOD}_{\text{AERONET}})$ (Sharma et al. 2021). In this study, MODIS Aqua (from 2002 to 2022) and Terra (from 2001 to 2022) collection 6.1 (C6.1) AOD data from the Kanpur region are utilized, and AOD is retrieved via the DT and DB algorithms. Table 1 depicts the scientific dataset set (SDS) used in the current work. The current study downloads the $DT_{3 \text{ km}}$, $DT_{10 \text{ km}}$, and $DB_{10 \text{ km}}$ AOD products from <https://ladsweb.modaps.eosdis.nasa.gov/>.

2.2.3. Meteorological data

The ERA5 meteorological dataset, generated by ECMWF, is a global reanalysis that offers hourly estimates of atmospheric, terrestrial, and oceanic climate variables from 1940 to the present (Hersbach et al. 2020). With its high temporal and spatial resolution, ERA5 is widely used to evaluate AOD changes during dust storms, biomass burning, and seasonal variations. Meteorological parameters such as temperature, humidity, surface pressure, wind speed, and wind direction play a pivotal role in comprehending AOD dynamics due to their influence on aerosol distribution, chemical composition, and optical characteristics. Pressure gradients influence aerosol mixing, whereas humidity and temperature impact hygroscopic growth. Wind parameters contribute to the analyses of aerosol transport and dispersion (Jiang et al. 2021). Additionally, radiation and heat fluxes

Table 1. Details of the MODIS aerosol products.

Satellite	Data	Scientific Data Set (SDS)	Description	Resolution
Terra	MOD04_L2	Optical_Depth_Land_And_Ocean	DT AOD at 550 nm over land and ocean	10 km
		Deep_Blue_Aerosol_Optical_Depth_550_Land_Best_Estimates	DB AOD at 550 nm over land	
	MOD04_3K	Optical_Depth_Land_And_Ocean	DT AOD at 550 nm over land and ocean	3 km
Aqua	MYD04_L2	Optical_Depth_Land_And_Ocean	DT AOD at 550 nm over land and ocean	10 km
		Deep_Blue_Aerosol_Optical_Depth_550_Land_Best_Estimates	DB AOD at 550 nm over land	
	MYD04_3K	Optical_Depth_Land_And_Ocean	DT AOD at 550 nm over land and ocean	3 km

resulting from solar and thermal energy at the surface and the Top of Atmosphere (TOA) significantly influence atmospheric and surface conditions, affecting aerosol concentration and distribution.

Incorporating these factors improves the precision of AOD prediction models by accounting for fundamental environmental interactions. This work uses various meteorological parameters, including U and V wind components, air temperature, dew point temperature, skin temperature, surface pressure, boundary layer height, relative humidity, heat flux, solar, thermal, and UV radiation data, to correct MODIS AOD along with AERONET AOD. The study does not directly use the U and V wind component data. Instead, the wind speed (WS) and wind direction (WD) are determined based on this data and incorporated into the dataset. Equations (3) and (4) give the formula for calculating wind speed and direction (ECMWF 2024).

$$\text{Wind Speed} = \sqrt{u^2 + v^2} \quad (3)$$

$$\text{Wind direction} = \text{mod}\left(180 + \left(\frac{180}{\pi}\right)\text{atan2}(u, v), 360\right) \quad (4)$$

In addition, the study includes time parameters such as year, DOY, Modified DOY (MDOY), and MDOY direction, as suggested by Lanzaco et al. (2016). The inclusion of a time parameter serves to capture both long-term and seasonal variations in AOD.

It should be noted that in this study, we used the ERA5 hourly data on single levels for all the meteorological parameters except relative humidity. Relative humidity was instead obtained from the ERA5 hourly data on pressure levels, specifically at the 950 hPa level. This level lies within the lower troposphere and planetary boundary layer, where aerosols most strongly interact with moisture, affecting their hygroscopic growth and AOD variability (Cao et al. 2021). A detailed description of all these parameters is provided in Table 2, and the workflow of this study is presented in the next section.

3. Methodology

The complete workflow of this paper is shown in Figure 2. It starts with dataset collection, then pre-processing the data for spatiotemporal matching, finding the relevant data by feature selection methods, splitting the matched data into training and test parts, and

Table 2. Summary of the dataset downloaded.

Type	Source	Parameters	Spatial Resolution	Temporal Resolution
Ground truth data	AERONET	Aerosol Optical Depth	440 nm, 500 nm, and 675 nm	15 min
Satellite data	MODIS (Terra and Aqua)	Dark Target AOD at 3 km Dark Target AOD at 10 km Deep Blue AOD at 10 km	3 km 10 km 10 km	1 day
Meteorological data	ECMWF-ERA5	10 m u component of wind (u10) 10 m v component of wind (v10) 2 m dew point temperature (d2m) 2 m temperature (t2m) Skin temperature (skt) Relative humidity at 950 hPa (R_950) Surface pressure (sp) Surface latent heat flux (slhf) Boundary layer height (blh) Top net thermal radiation (ttr) Top net solar radiation (tsr) Surface net thermal radiation (str) Surface net solar radiation (ssr) Downward UV radiation at the surface (uvb)	0.25° x 0.25°	1 hour
Time	–	Year Day of the year (DOY) Modified DOY (MDOY) MDOY direction (MDOY D)	–	–

applying a standard scalar. After that, a suitable ML/DL model is selected, the hyperparameters are tuned to find the generalized model using cross-validation, and the selected ML/DL model is trained with the best parameters to correct the MODIS AOD. Finally, various performance metrics are evaluated for the utilized models and compared to find the best model for AOD prediction. This work uses Python programming to implement all the models. This methodology is applied to six different MODIS datasets, such as Terra ($DT_{3\text{ km}}$, $DT_{10\text{ km}}$ and $DB_{10\text{ km}}$) and Aqua ($DT_{3\text{ km}}$, $DT_{10\text{ km}}$ and $DB_{10\text{ km}}$). The subsequent sections briefly explain the methodology.

3.1. Dataset preparation

3.1.1. Data acquisition and spatiotemporal collocation

This paper considers three data sets over the Kanpur region: AERONET AOD, MODIS AOD and the meteorological data required for the MODIS AOD bias correction. All three data sets must undergo spatio-temporal matching before the pre-processing step. The version 3 level 2 AOD data can be downloaded from the AERONET website. It provides AOD data at 500 nm. As the MODIS AOD data is available at 550 nm, AERONET AOD is interpolated from 500 to 550 nm using the formula in Section 2.2.1. The LAADS DAAC website offers access to the MODIS Collection 6.1 daily AOD data. This study uses nearly two decades of AOD observations from the Terra satellite (2001–2022) and Aqua satellite (2002–2022). It focuses on MODIS Terra/Aqua aerosol products, specifically MOD04C6.1/MYD04C6.1, with a quality assurance flag (QAF) of 3. These products provide AOD measurements at 10 km and 3 km spatial resolutions. The DT method retrieves AOD at both resolutions, while the DB algorithm retrieves AOD only at 10 km. The downloaded MODIS data must be spatially

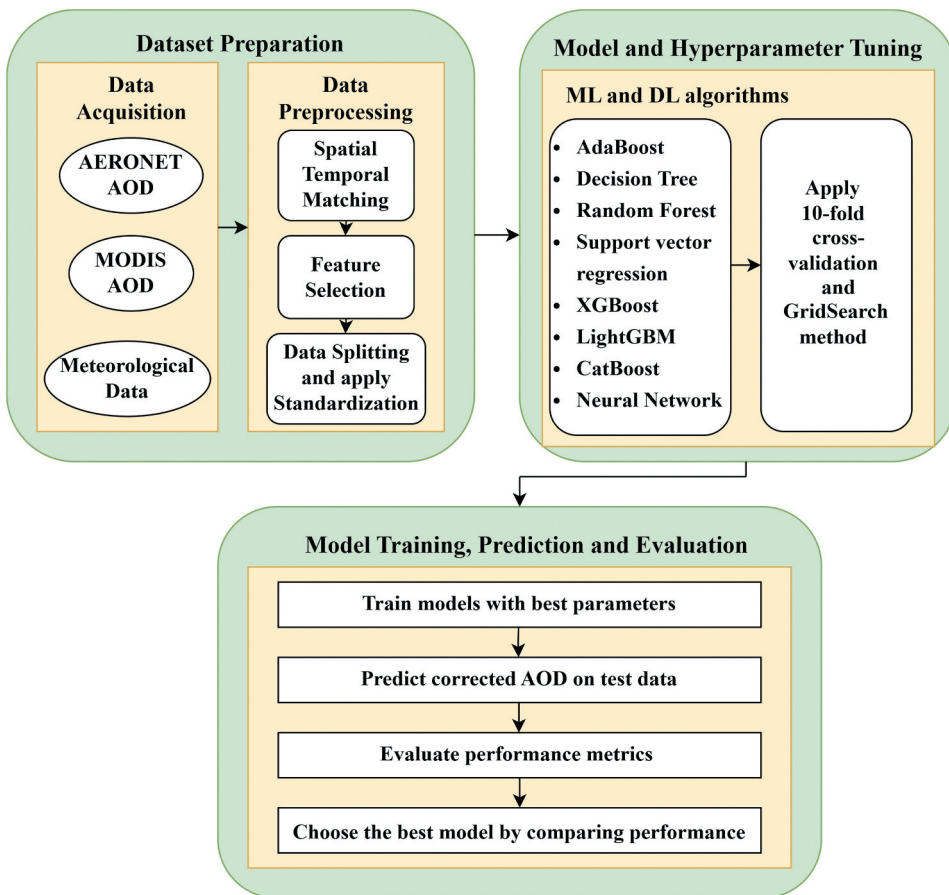


Figure 2. Flowchart of the methodology.

and temporally aligned with corresponding AERONET site data. A sampling window with 3×3 pixels centred on the AERONET station is used to compute the MODIS AOD values for each daily file. For AERONET AOD, a time window of ± 30 minutes around the MODIS (Terra and Aqua) overpass time is used to select daily measurements. AERONET data is considered only if there are at least two AOD retrievals (Choudhry, Misra, and Tripathi 2012). After collecting valid AOD data, the daily averages are calculated, and files that do not meet the collocation criteria are excluded from the analysis.

Kanpur AERONET site coordinates are provided for the ERA5 data download, and the downloaded meteorological parameters must be collocated over time with AERONET AOD and MODIS AOD. The correspondence between the MODIS passing time and the meteorological data is essential because hourly meteorological data is only accessible from the ECMWF. As suggested by Lanzaco et al. (2016), if the MODIS pass occurs within the first 30 minutes of an hour, the meteorological data from that same hour is used. However, if the pass occurs after 30 minutes, the data from the next hour is used. This approach ensures that the selected meteorological information closely represents the actual atmospheric conditions during the MODIS overpass. After the spatiotemporal collocations, the total data points of 2663, 2607, 2837, 2965, 2929 and 3113 are available

for model training and testing, respectively, for Aqua $DT_{3\text{ km}}$, Aqua $DT_{10\text{ km}}$, Aqua $DB_{10\text{ km}}$, Terra $DT_{3\text{ km}}$, Terra $DT_{10\text{ km}}$ and Terra $DB_{10\text{ km}}$. Before the model training, these data must be split into train and test sets to predict the corrected MODIS AOD.

3.1.2. Feature selection

This study uses multiple ML algorithms to assess the importance of features and understand their contributions to the predictive models. Since various ML algorithms are used for AOD prediction, the significance of the feature drawn from a single algorithm is not enough to select the best features suitable for all the models. Feature importance extracted using various methods ensures comprehensive and robust insights. Traditional feature importance was derived from ensemble models, including RF, AdaB, DT, XGB, CatB and LGBM, which rank the features based on their contribution to impurity reduction in the splits during training. It is a straightforward way to identify which features most influence model predictions. Secondly, permutation importance was used with RF, AdaB, DT, XGB, CatB, LGBM and SVR. This method finds the significance of each feature by computing the decrease in model performance when the feature values are shuffled randomly. It reflects the impact of feature values on model predictions, independent of the model's inherent structure (Nirmalraj et al. 2023). SHAP values were calculated for RF, DT, XGB, CatB and LGBM models to complement these techniques. Based on game theory, it assigns each feature a value. Features with higher SHAP values have a more significant influence on the model's predictions (H. Wang et al. 2024).

Since this work utilizes multiple methods for identifying the best features, we adopted a rank-based feature aggregation method. In this approach, feature importance is first computed separately for each method. Within each method, features were ranked so that the most crucial feature received rank 1, the next most important received rank 2, and so on. After ranking features within each method, the average rank across all methods was calculated for every feature. A feature with a smaller mean rank value is considered more important, as it consistently appears near the top across multiple methods. These aggregated average ranks were then used to identify the most influential features. This approach ensures feature importance scores from different methods are normalized and comparable. The rank-based feature aggregation technique was used for both Aqua and Terra satellites, and the outcomes are shown in Figure 3(a,b) as bar graphs for Aqua and Terra, respectively. In these figures, the features at the bottom of the bar chart correspond to higher importance (smaller mean rank values), whereas those at the top correspond to lower importance. The 15 parameters from both charts are selected as the best features, including 3 MODIS AOD, meteorological data and time parameters. By integrating these methods – feature importance, permutation importance, and SHAP, 13 essential features are selected. They are MODIS AOD, wind speed, wind direction, t2m, d2m, R_950, slhf, blh, sp, year, DOY, ttr and str to train the ML models for predicting corrected MODIS AOD for each dataset.

3.1.3. Data splitting and standardization

Machine learning develops algorithms to learn from behaviours or properties of datasets. The dataset needs to be separated into two sets: train and test. The training dataset is used to train the selected models, and the trained models can predict test set data. This work uses the split ratio of 80:20 for the training and testing process for all the ML

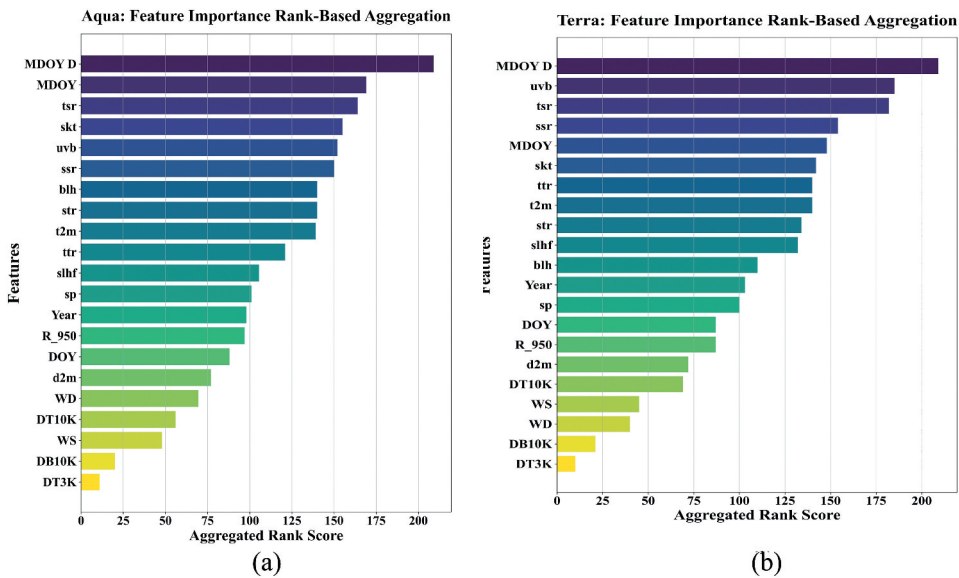


Figure 3. Feature importance for (a) Aqua satellite and (b) Terra satellite.

algorithms. ANN needs training, validation, and test datasets. To make the comparison consistent for all the models, we selected the same 20% of data for testing, and the remaining 80% of data can be split into a ratio of 80:20 for training and validation, respectively. The test dataset size details are given as follows: Aqua $DT_{3\text{ km}}$ – 533, Aqua $DT_{10\text{ km}}$ – 522, Aqua $DB_{10\text{ km}}$ – 568, Terra $DT_{3\text{ km}}$ – 593, Terra $DT_{10\text{ km}}$ – 586 and Terra $DB_{10\text{ km}}$ – 623.

The features in real-world datasets often have different scales, leading to issues in ML and DL models. Algorithms may favour features with more extensive ranges, even if they aren't the most informative. Scaling is thus one of many pre-processing steps that must be undertaken before applying a ML model to a dataset. It standardizes the feature ranges, ensuring equal contribution to model learning. Standard scalar is a compelling method that implements Z-score normalization by transforming features to have a mean of zero and a standard deviation of one, balancing the impact of both positive and negative values. The formula for the standard scalar is provided in Equation (5) (De Amorim, Cavalcanti, and Cruz 2023). By using the μ - mean and s - standard deviation, the x_i is transformed to x'_i .

$$x'_i = \frac{x_i - \mu}{s} \quad (5)$$

3.2. ML and DL algorithms

Generally, traditional supervised ML regressors are trained with labelled data and are used for many remote sensing applications. This work uses multiple ML algorithms, including RF, AdaB, DTree, XGB, CatB, LGBM, SVR, and ANN, to correct the bias generated by the MODIS satellite retrieved AOD. The best hyperparameters for all the models are selected using GridSearchCV and 10-fold cross-validation. Since cross-validation is used, the

standard scalar and regressors are defined in the pipeline to ensure the proper data scaling. A description of all the algorithms is given next.

3.2.1. AdaBoost regressor

Adaptive Boosting (AdaBoost) is a supervised ensemble learning algorithm that trains weak learners (an algorithm that achieves a performance slightly better than random guessing) sequentially, typically DTrees, to enhance prediction accuracy for classification and regression problems. Unlike RF, which trains learners simultaneously, it adds weak learners iteratively, each improving on the previous one's errors (Gao et al. 2010). All training samples initially have the same weights, but incorrectly classified instances get greater weights in successive iterations. In regression, weak learners are averaged through weighted averages instead of voting. This iterated optimization converges on model error exponentially, with increasing predictive precision aimed at challenging instances. The hyperparameters, such as n estimators (number of DTrees), learning rate, and loss functions, are fine-tuned for an accurate prediction.

3.2.2. Decision Tree regressor

A Decision Tree is a supervised ML algorithm that recursively partitions the data based on measures such as the Gini index or Information Gain to construct a tree-based model for making predictions (Rajendiran and Kumar 2023). It splits until it has pure leaf nodes, where, in classification, a majority vote is used, but in regression, it predicts the average value within a leaf node. DTree can be implemented with ID3, C4.5, or CART (Classification And Regression Tree), where CART is more suited to numerical data because it supports handling outliers, missing values, and cost-complexity pruning. CART is utilized with the Gini criterion in this research for regression problems, using post-pruning to avoid overfitting and improve generalization. The DTree performance can be improved by adjusting the hyperparameters like max depth (maximum depth of the tree), max features (number of features considered for best split), min samples split (minimum number of samples required to split an internal node), and min samples leaf (minimum number of samples needed to be at a leaf node).

3.2.3. Random forest regressor

Random Forest is an ensemble algorithm that addresses regression and classification issues by aggregating various DTrees with the bagging (bootstrapping + aggregation) method (Mohan, Manisekaran, and Kumar 2021). It generates varied training sets and chooses random subsets of features for every DTree to provide independent training and enhance model strength. In regression, RF produces numerical predictions as an average output of all the DTrees, thereby avoiding overfitting while enhancing accuracy from an individual DTree. Though categorical and numeric data can be processed, interpreting RF is slightly more complex than interpreting an individual DTree. The key hyperparameters include the n estimators, max features, max depth, min samples split, and min samples leaf, which are tuned to get the generalized RF model.

3.2.4. Support vector regressor

SVM is a supervised ML algorithm for classification (SVC) and regression (SVR) that identifies the best hyperplane to classify data and maximize the margin between support

vectors (Lary et al. 2009). For non-linearly separable data, kernel functions such as Radial Basis Function (RBF) transform data into a higher space where a linear boundary can be defined. SVR hyperparameters, such as the type of kernel (e.g. linear, RBF, poly, and sigmoid), regularization parameter (C), and gamma, are adjusted for best model performance in regression.

3.2.5. XGBoost regressor

XGBoost (eXtreme Gradient Boosting) is a classification and regression ensemble ML algorithm that employs sequential boosting, in which every DTree corrects the residual errors of the previous one with the help of gradient descent. It utilizes L1 (alpha) and L2 (lambda) regularization to avoid overfitting and makes second-order gradient computation, cache optimization, and parallelization to improve efficiency. XGBoost suits highly complex datasets and enhances prediction accuracy (Rajendiran and Kumar 2023). Various hyperparameters such as maximum depth, learning rate, the number of trees, minimum child weight, gamma function (minimum loss reduction required for a split), subsample (fraction of samples used for training each tree) and colsample bytree (fraction of features used for training each tree) are fine-tuned using GridSearchCV to improve model performance.

3.2.6. LightGBM regressor

The Light Gradient Boosting algorithm is a boosting framework based on gradients optimized for performance and extensive data handling. LightGBM adds Gradient-Based One-Side Sampling (GOSS) to keep essential samples and Exclusive Feature Bundling (EFB) to bundle mutually exclusive features (Tian et al. 2021). Unlike traditional boosting, it follows a leaf-wise growth strategy, improving accuracy but increasing overfitting risk, which is mitigated by L1 and L2 regularization. LightGBM is quicker than XGBoost and optimized for scalable ML applications (Rajendiran, Sebastian, and Kumar 2024). The hyperparameters like the number of leaves, maximum depth, n estimators, learning rate, subsample and colsample bytree are adjusted to enhance its performance.

3.2.7. CatBoost regressor

CatBoost is a symmetric tree-structured gradient-boosting algorithm with computational efficiency. It differs from XGBoost and LightGBM as it applies ordered boosting to mitigate overfitting and target leakage using permutation-based training (Jabeur et al. 2021). It processes categorical features natively without complex pre-processing and supports multiple data types. It is highly efficient in handling structured data, preventing overfitting and making accurate continuous predictions (Rajendiran, Sebastian, and Kumar 2024). Various hyperparameters, such as iterations (number of trees), learning rate, depth (maximum depth of the tree), and L2 leaf reg (L2 regularization for leaf weights), are tuned finely to get the optimal model for MODIS AOD prediction.

3.2.8. Artificial neural network regressor

For the deep learning component of this study, we employed an ANN, also known as a Multi-Layer Perceptron (MLP). This architecture was chosen because the problem is a regression task and does not require spatial or sequential modelling, as would be the case with convolutional neural networks (CNNs) or recurrent neural networks (RNNs). The

ANN was optimized using GridSearchCV with 10-fold cross-validation. Artificial Neural Networks are nonparametric ML models inspired by biological neural systems. They are widely used for function approximation, classification and correcting biases across various fields, including geoscience, oceanography, and remote sensing. ANNs consist of interconnected processing units called neurons arranged in input, hidden, and output layers. It is also named as multi-layer perceptron since it has multiple layers. Each neuron computes its inputs through weighted connections, with a transfer function that produces its outputs. The k^{th} neuron's output can be represented as the weighted sum of the inputs, as shown in Equation (6).

$$y_k = \varphi \left(\sum_{j=1}^n w_{jk} x_j \right) \quad (6)$$

where x_j stands for the n input variables to the neuron, w_{jk} for the weight from unit j to k and φ is the transfer function. ANN learns from labelled input-output pairs. During training, the given data is divided into training, validation, and test sets, and the splitting details can be found in Section 3.1.3. However, the idea is to modify the weights and biases of the network to minimize errors such as RMSE between predicted outputs and actual targets. The training set fine-tunes the model's parameters, the validation set tracks convergence and prevents overfitting, and the test set analyses the final model's generalization performance. Then, training goes on in epochs. At each epoch, RMSE is computed to better or best the network weights until predictions approximate the values of the corresponding targets (Malakar et al. 2012). ANNs can systematically capture complicated, nonlinear input-output relationships and adapt well to various datasets without making strong assumptions.

The implementation of ANN uses TensorFlow with the Keras library. Like ML algorithms, the best hyperparameters of ANN are selected using the GridSearchCV along with 10-fold cross-validation. The parameters, including the number of hidden layers, number of neurons for each hidden layer, activation function, learning rate, L2 regularization values, and epochs, are given in the parameter grid to find the best sets for each dataset. We also selected the optimizer as Adam, batch size as 32, and dropout rate as 0.2 for training the ANN model. ANN has input, hidden and output layers, and Figure 4 illustrates the architecture of ANN. Features are given as input to the input layer; we have 13 features in our work. Similarly, we have one output layer since MODIS AOD prediction is a regression problem, and this layer produces continuous outputs. For the six data sets, the GridSearchCV gives the number of hidden layers as three during the hyperparameter tuning process, so we had hidden layer 1, hidden layer 2 and hidden layer 3. The L2 regularization technique is applied to all the hidden layers to avoid overfitting. After each hidden layer, we provide Batch normalization, Activation, and dropout layers to get a more generalized model. The purpose of each layer is explained next.

Batch normalization regularizes neural networks by minimizing internal covariate shifts, accelerating convergence and stabilizing training (Lofte and Szegedy 2015). By incorporating batch normalization, feature maps are normalized to have unit variance and zero mean, accelerating training and minimizing distribution changes in intermediate layers. Activation functions provide nonlinearity, allowing networks to learn intricate patterns. Essential activation functions are ReLU (Rectified Linear Unit – fast but can lead to dead neurons), Leaky ReLU (permits tiny gradients for negative inputs), ELU

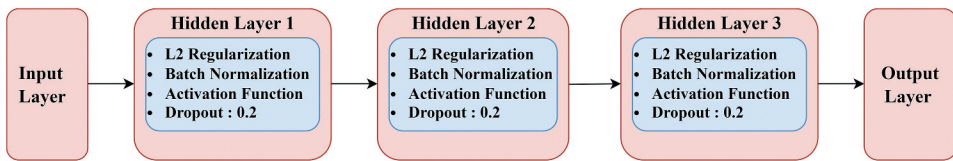


Figure 4. ANN architecture.

(Exponential Linear Unit – smoothly decreasing vanishing gradients), and Swish (non-monotonic, enhancing deep network performance) (Dubey, Singh, and Chaudhuri 2022).

This work includes four activation functions, ReLU, Leaky ReLU, ELU and Swish, to ensure the best hyperparameter selection process. Those functions provide better insight into the model performance compared with other activations. Since it is a regression problem, we have selected a linear activation function for the output layer. A dropout layer is a regularization technique in neural networks where some fractions of the neurons' outputs are randomly set to zero during training. These neurons do not contribute to the output computation of the network for specific training iterations. At the same time, the model learns more robust and generalized patterns rather than focusing too much on particular neurons. In addition to this, we have used one more technique called step decay. It is a learning rate scheduling method used during training to decrease the learning rate gradually. It facilitates smaller, more accurate weight updates by lowering the learning rate at the end stages of training; hence, convergence to a more optimal solution is achieved. Thus, step decay maintains a balance between speed and precision during training to avoid overshooting the optimal solution or stagnation at a suboptimal point. Adapting all these techniques into the ANN is an optimal model for correcting MODIS AOD based on meteorological variables and AERONET AOD as ground truth.

3.2.9. Cross-validation and hyperparameter tuning

In this paper, 10-fold cross-validation is used to improve training efficiency. In this, the dataset is split into ten equal subgroups. In each iteration, one subgroup is kept aside as the test set, and the remaining nine groups are taken as the training data. The above process is repeated 10 times to ensure that every subgroup becomes the test set exactly once. This method eliminates bias since most data are utilized in the model's training while having robust cross-validation at each iteration. The ANN model's weights are updated at each iteration in training, improving learning capability. 20% of the test data is kept aside for validation, and 80% of the training data is used for the 10-fold cross-validation. An automated optimization approach, such as a grid search for hyperparameter tuning, is used. It systematically evaluates the predetermined hyperparameters to determine the best configuration for the model. This eliminates the old trial-and-error approach, thereby picking the optimum parameters for maximizing the model's performance. With 10-fold cross-validation and a grid search, the model's size is minimized, with reduced overfitting and enhanced generalization capability. Details of the best hyperparameters for each model over six datasets are provided in Table S1 of the supplementary material.

3.3. Model training, prediction, and evaluation

After that, all the ML and DL models were trained with the best hyperparameters obtained for correcting MODIS Terra and Aqua AOD over three retrieval algorithms using meteorological variables and AERONET AOD. To evaluate the effectiveness of each model, various performance metrics are adopted in this work in addition to the expected error boundary. The model performance has been assessed using the Pearson correlation coefficient (R), RMSE, MAE, MB, MAPE, and Expected Error (EE). Equations (7) - (14) define the performance metrics.

$$R = \frac{\sum_{i=1}^n \left((AOD)_{act,i} - \overline{(AOD)}_{act} \right) \left((AOD)_{pred,i} - \overline{(AOD)}_{pred} \right)}{\sqrt{\sum_{i=1}^n \left((AOD)_{act,i} - \overline{(AOD)}_{act} \right)^2} \sqrt{\sum_{i=1}^n \left((AOD)_{pred,i} - \overline{(AOD)}_{pred} \right)^2}} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left((AOD)_{act,i} - (AOD)_{pred,i} \right)^2} \quad (8)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |(AOD)_{act,i} - (AOD)_{pred,i}| \quad (9)$$

$$MB = \frac{1}{N} \sum_{i=1}^N \left((AOD)_{pred,i} - (AOD)_{act,i} \right) \quad (10)$$

$$MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{(AOD)_{act,i} - (AOD)_{pred,i}}{(AOD)_{act,i}} \right| \quad (11)$$

$$EE_{DT10K} = \pm (0.05 + 0.15 \times (AOD)_{AERONET}) \quad (12)$$

$$EE_{DT3K} = \pm (0.05 + 0.20 \times (AOD)_{AERONET}) \quad (13)$$

$$EE_{DB10K} = \pm (0.03 + 0.20 \times (AOD)_{AERONET}) \quad (14)$$

After that, the trained models use the unseen test data to predict the corrected satellite AOD. The predicted AOD will then be validated against the AERONET AOD for each model and each MODIS retrieval algorithm. The performance of various models is assessed to find the best model on each dataset by utilizing the above-mentioned error metrics, correlation, and percentage of data within the expected error boundary. Three main types of EE are distinguished: i) above EE, which denotes overestimation; ii) within EE, which denotes correct estimation; and iii) below EE, which denotes underestimation. The obtained results and related explanations are provided in the next section.

4. Results and discussion

This section describes the validation results of ML and ANN predicted AOD against AERONET AOD for six datasets. There are three different MODIS retrieval algorithms for Aqua and Terra satellites, including $DT_{3 \text{ km}}$, $DT_{10 \text{ km}}$, and $DB_{10 \text{ km}}$. Each dataset includes any

of the MODIS AOD, along with meteorological and time parameters as features for the models to predict corrected AOD. In this Section, the model output for each dataset is discussed separately. Before that, the validation comparison of all the retrieval algorithms is presented with total data points to analyse how each MODIS algorithm performs retrieving AOD at the Kanpur site. In addition, the performance metric-wise comparison of all the datasets and models was given at the end. Following this, the computational efficiency analysis of the ML and ANN models, in terms of runtime and memory usage, is also provided.

4.1. Validation of MODIS Terra and Aqua aerosol products against AERONET

Figure 5 depicts the evaluation results of the MODIS Aqua and Terra satellite retrieved AOD using the $DT_{3\text{ km}}$, $DT_{10\text{ km}}$ and $DB_{10\text{ km}}$ algorithms against AERONET-derived AOD over Kanpur during the entire study period (Terra: 2001–2022 and Aqua: 2002 – 2022). The regression plots in Figure 5(a–c) are for the MODIS Aqua satellite. In contrast, the regression plots between AOD_{AERONET} and AOD_{MODIS} data for the MODIS Terra satellite are shown in Figure 5(d–f), respectively, for $DT_{3\text{ km}}$, $DT_{10\text{ km}}$, and $DB_{10\text{ km}}$ AOD products. Key statistical metrics, such as R, RMSE, MAE, MAPE, MB, WEE, AEE and BEE, are analysed to evaluate the performance of these algorithms.

All the figures use density to better present data distribution by representing concentrations of data points using colour gradients, where red represents a greater density, and blue indicates a smaller density. This method also makes it possible to find data points where AERONET and MODIS AOD agree better. The highly dense red spots near the 1:1 line indicate strong consistency between the two datasets, underscoring the algorithm's correctness and effectiveness. It also captures overestimation or underestimation trends and identifies outliers in low-density regions. Density is computed using Kernel Density Estimation (KDE), which calculates the density of data points in two-dimensional space and normalizes it across each subplot for consistent interpretation. This overlay of density on scatter plots provides a simultaneous view of individual points and the behaviour of points collectively, which helps to better understand algorithm performance and where improvements are necessary. For example, Terra's $DT_{3\text{ km}}$ algorithm has areas of high reliability near the 1:1 line and consistent AOD retrievals compared to other figures.

The analysis of Figure 5 reveals that the $DT_{3\text{ km}}$ algorithm performs well for Aqua and Terra satellites due to its design for higher spatial resolution aerosol retrievals, which are particularly effective for monitoring aerosol properties at finer scales. For Aqua, it achieves a strong correlation ($R = 0.729$) with AERONET AOD and 66.95% of data falling within MODIS EE bounds. However, it produces slightly higher error values (RMSE = 0.285, MAE = 0.172, and MAPE = 28.24%). Also, it shows a slight overestimation tendency, with 24.11% of the data exceeding the upper EE boundary. In contrast, the Aqua $DT_{10\text{ km}}$ algorithm has a reduced correlation ($R = 0.713$) and lower errors (RMSE = 0.267, MAE = 0.171, and MAPE = 28.10%), with only 58.38% of the data within EE bounds and a notable overestimation trend (27.66% above EE). The $DB_{10\text{ km}}$ algorithm for Aqua performs slightly better than the $DT_{10\text{ km}}$, with an R of 0.732, along with reduced error values (RMSE = 0.271, MAE = 0.166, and MAPE = 25.91%) and 63.66% of data within EE bounds with high underestimation (25.48%). For Terra, the $DT_{3\text{ km}}$ algorithm demonstrates the highest reliability, with the strongest correlation ($R = 0.796$), slightly higher errors (RMSE = 0.263, MAE = 0.160, MAPE

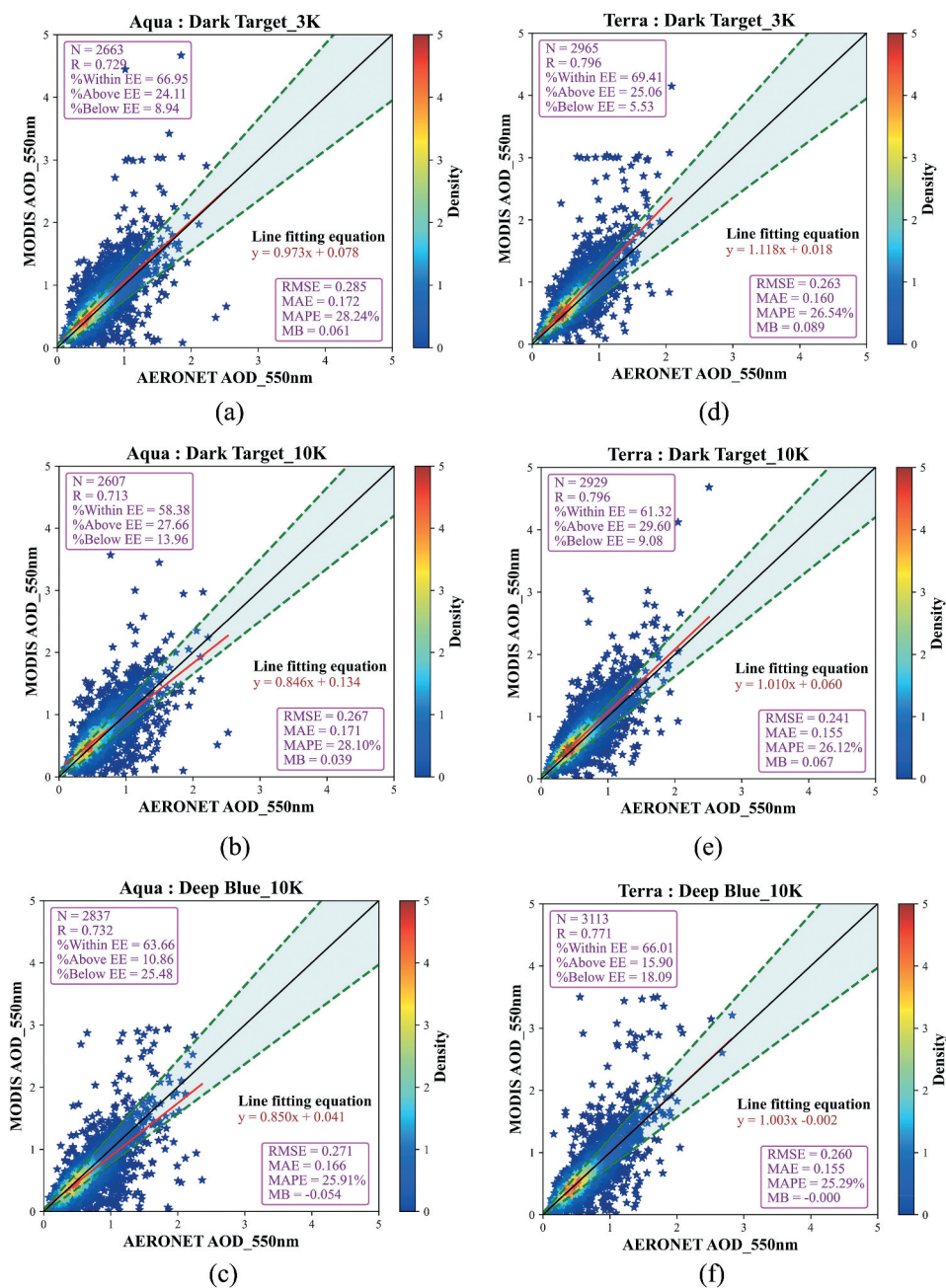


Figure 5. Validation results of MODIS retrieved AOD against AERONET AOD over Kanpur for both Aqua (a, b, c) and Terra (d, e, f) satellites, where (a, d) – $DT_{3\text{ km}}$, (b, e) – $DT_{10\text{ km}}$, and (c, f) – $DB_{10\text{ km}}$. The solid black line is a 1:1 line, the dashed green lines with light blue shades indicate the expected error boundary, the red line is a regression line, and the blue stars are data points.

= 26.54%), and a higher percentage (69.41%) of the data within EE bounds, with minimal underestimation (5.53%) as compared to the other two algorithms. The $DT_{10\text{ km}}$ algorithm for Terra performs slightly worse, with 61.32% of data within EE bounds, showing higher overestimations (29.60%). Meanwhile, the $DB_{10\text{ km}}$ algorithm for Terra demonstrates good performance, with 66.01% of data within EE bounds and minimal overestimations compared to other algorithms. Although it has slightly lower R (0.771) and error values than the $DT_{3\text{ km}}$, it remains a strong alternative due to its balanced data estimation.

The least-squares regression fit's intercept and slope values reveal biases in the two measurement techniques. It is discovered that all aerosol retrieval techniques resulted in regressions with slopes greater than zero, suggesting a systematic positive bias in AOD_{MODIS} as $AOD_{AERONET}$ rises. However, for the Aqua satellite's AOD estimation, the slope magnitude varies from 0.85 to 1, indicating a tendency towards slight underestimation. In contrast, the Terra satellite's slope value exceeds 1, suggesting a slight overestimation. The intercept values of the Aqua satellite are higher than those of the Terra satellite for Figure 5. Another parameter called MB also verifies the underestimation and overestimation of AOD_{MODIS} , where the negative value denotes the underestimation, and the positive value indicates the overestimation. For both Terra and Aqua satellites, the MB value is positive for $DT_{3\text{ km}}$ and $DT_{10\text{ km}}$, representing the higher overestimation of AOD_{MODIS} .

In contrast, the $DB_{10\text{ km}}$ plots of both satellites show negative MB values, indicating more underestimation of MODIS AOD. In summary, the $DT_{3\text{ km}}$ emerges as the most robust option for both Aqua and Terra satellites based on the data within the EE, with Terra consistently outperforming Aqua in terms of correlation, error metrics, and the percentage of data within EE bounds. The $DT_{10\text{ km}}$ and $DB_{10\text{ km}}$ algorithms for Aqua and Terra also perform well, particularly in minimizing the error metrics compared to $DT_{3\text{ km}}$. These findings underscore the importance of algorithm selection for accurate AOD retrieval and highlight opportunities for further optimization to reduce biases and improve reliability.

4.2. Performance comparison of ML and ANN models for MODIS Aqua $DT_{3\text{ km}}$

Figure 6 illustrates the validation results of multiple ML and ANN models used to correct biases in AOD retrieved from the MODIS Aqua $DT_{3\text{ km}}$ algorithm. The models are evaluated against AERONET AOD observations, which serve as the reference dataset. There are nine subplots, and each plot represents a different model. The Figures 6(a–i) are intended for the validation of MODIS AOD, AdaB AOD, DTree AOD, RF AOD, SVR AOD, XGB AOD, LGBM AOD, CatB AOD and ANN AOD against AERONET AOD, respectively. The exact order is used for the outputs of the remaining datasets. The goal is to compare their performance in predicting AOD values using a test dataset that aligns closely with the AERONET ground-based observations. The scatter plots show the relationship between the AERONET AOD values on the x-axis and the predicted AOD values from the respective model on the y-axis. A 1:1 line (solid black) is included in each plot as a benchmark for perfect agreement between the predicted and reference values, while the dashed green lines with light blue colour shades represent error margins. The density of data points is indicated by the colour bar on the right, where warmer colours represent higher densities. The comparison involves vital performance metrics such as R, RMSE, MAE, MAPE, MB and

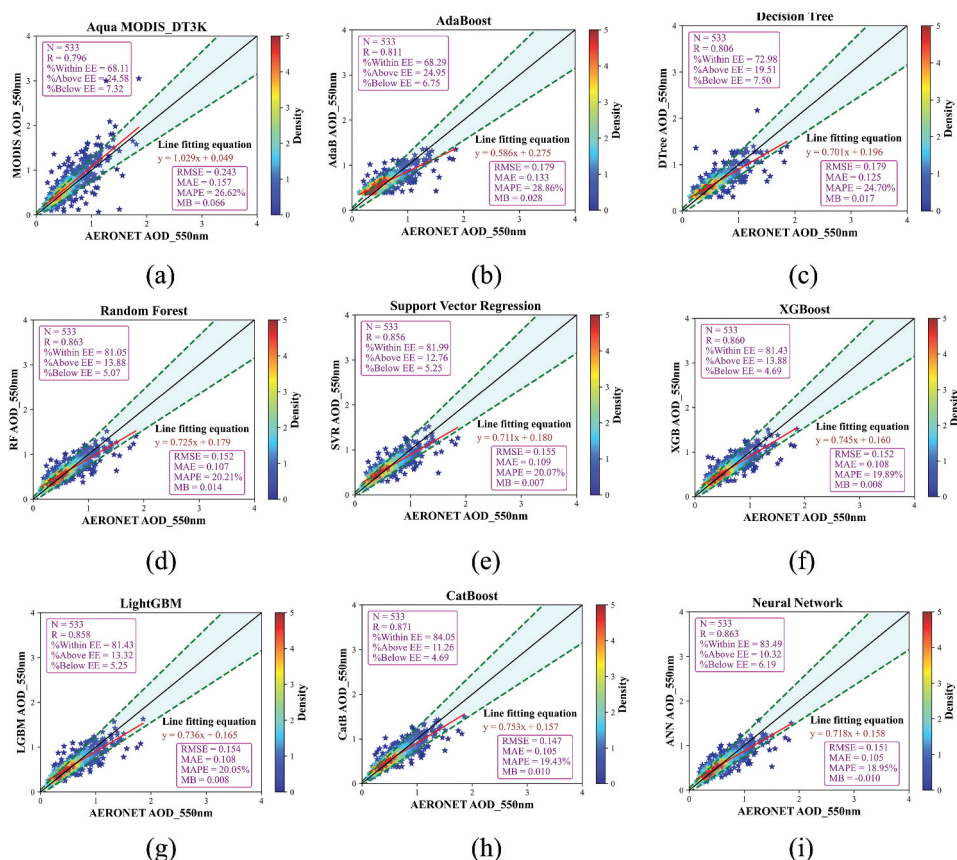


Figure 6. Validation of ML and ANN models predicted AOD against AERONET AOD for MODIS Aqua $DT_{3\text{ km}}$.

EE. These metrics highlight how well each model aligns MODIS AOD predictions with AERONET observations. Figure 6(a) shows the validation of Aqua MODIS $DT_{3\text{ km}}$ retrieved AOD against AERONET AOD, which acts as the baseline model. This model results have an R of 0.796, RMSE of 0.243, MAE of 0.157 and MAPE of 26.62% with an MB of 0.066, indicating moderate performance in reproducing AERONET AOD values. The percentage of data within the EE is 68.11, with AEE and BEE values of 24.58% and 7.32%, respectively. By providing MODIS AOD and some meteorological parameters as features, the selected ML and ANN models correct the bias introduced by MODIS satellites.

In this, AdaB performs slightly better, with WEE of 68.29%, R of 0.811, RMSE of 0.179, and MAE of 0.133, but it exhibits a higher MAPE of 28.86% compared to the baseline model. The DTree model shows reduced accuracy, with an R of 0.806, RMSE of 0.179, and reduced error values, including MAE of 0.125, MAPE of 24.70%, along with a notable increase in WEE, i.e. 72.98% as compared to AdaB. Ensemble methods, such as RF and XGB, show significant improvements. RF achieves an R of 0.863, WEE of 81.05%, RMSE of 0.152, and MAE of 0.107, with a MAPE close to 20.21%. XGB delivers comparable performance, with an R of 0.860, WEE of 81.43%, RMSE of 0.152, and MAE of 0.108 while

maintaining minimal MAPE (19.89%). SVR achieves similar performance, with an R of 0.856, RMSE of 0.155, and MAE of 0.109, but its WEE (81.99%) and MAPE (20.07%) are slightly higher than XGB. Among the ML models, gradient-boosting techniques such as LGBM and CatB are particularly effective. LGBM shows strong results with a WEE of 81.43%, R of 0.858, RMSE of 0.154, MAE of 0.108, and MAPE of 20.05%. Compared to all the models, CatB gives superior performance with an R of 0.871, WEE of 84.05%, RMSE of 0.147, and MAE of 0.105, with a lower MAPE of 19.43%. Finally, the ANN performs well, with an R of 0.863, WEE of 83.49%, RMSE of 0.151, MAE of 0.105 and reduced MAPE of 18.95%, demonstrating the capability of deep learning to model complex relationships. For all the models, the AEE shows a higher percentage than the BEE, indicating a higher overestimation. It can also be verified by MB, which shows positive values for all the models except ANN. All the models showed positive slopes, and intercepts denote the positive bias. Based on comparison analysis, CatB exhibits superior performance, with higher R values (above 0.871), lower RMSE, MAE, and MAPE, and a higher percentage of EE data than the other models considered, followed by ANN. Hence, this paper recommends CatB and ANN models for MODIS AOD bias correction of Aqua $DT_{3\text{ km}}$ aerosol product. Similarly, validation outcomes of various ML algorithms for other remaining datasets like Aqua $DT_{10\text{ km}}$, Aqua $DB_{10\text{ km}}$, Terra $DT_{3\text{ km}}$, Terra $DT_{10\text{ km}}$ and Terra $DB_{10\text{ km}}$ are presented in Figures S1 to S5 of the supplementary material, respectively.

4.3. Performance metric-wise comparison of ML and ANN models

Figure 7(a) presents the R -correlation values for various ML models and the MODIS AOD product, comparing predicted AOD values against ground-truth AERONET AOD. The MODIS retrievals generally exhibit lower correlation values, specifically 0.796, 0.666, 0.718, 0.821, 0.778 and 0.795 for Aqua $DT_{3\text{ km}}$, Aqua $DT_{10\text{ km}}$, Aqua $DB_{10\text{ km}}$, Terra $DT_{3\text{ km}}$, Terra $DT_{10\text{ km}}$, and Terra $DB_{10\text{ km}}$, respectively. Compared to all the tested models, DTree and AdaB demonstrate the poorest performances in predicting AOD across all the MODIS aerosol retrieval algorithms, as indicated by their lower correlation values. For Aqua $DT_{3\text{ km}}$, CatB achieves the highest correlation (0.871), followed closely by ANN, RF, XGB, LGBM, and SVR. In contrast, for Aqua $DT_{10\text{ km}}$, LGBM exhibits the best performance with a correlation of 0.796, followed by XGB, CatB, RF, ANN, and SVR. In the Aqua $DB_{10\text{ km}}$ dataset, CatB leads again with a correlation of 0.844, followed by LGBM (0.837) and XGB (0.83), showcasing their superior performance. For the Terra $DT_{3\text{ km}}$ dataset, SVR achieves the highest correlation (0.875), followed by CatB, ANN, RF, and LGBM. For Terra $DT_{10\text{ km}}$, CatB again outperforms other models with a correlation of 0.849, followed by XGB, SVR, RF, LGBM and ANN. Similarly, in the Terra $DB_{10\text{ km}}$ dataset, CatB achieves the highest correlation (0.88), followed by ANN, LGBM, XGB, RF and SVR. Overall, CatB consistently demonstrates superior performance in terms of correlation across multiple datasets, with the highest R -values observed in most cases. Overall, Terra $DB_{10\text{ km}}$ shows the highest correlation of 0.88 from CatB prediction compared to other algorithms. This highlights the robustness and effectiveness of the CatB algorithm in predicting AOD compared to other ML models.

Figure 7(b) focuses on the percentage of data points within the EE, which evaluates the proportion of predictions that meet a certain accuracy threshold. The MODIS AOD product lags, with the lowest percentage of data within EE, observed for Aqua $DT_{10\text{ km}}$

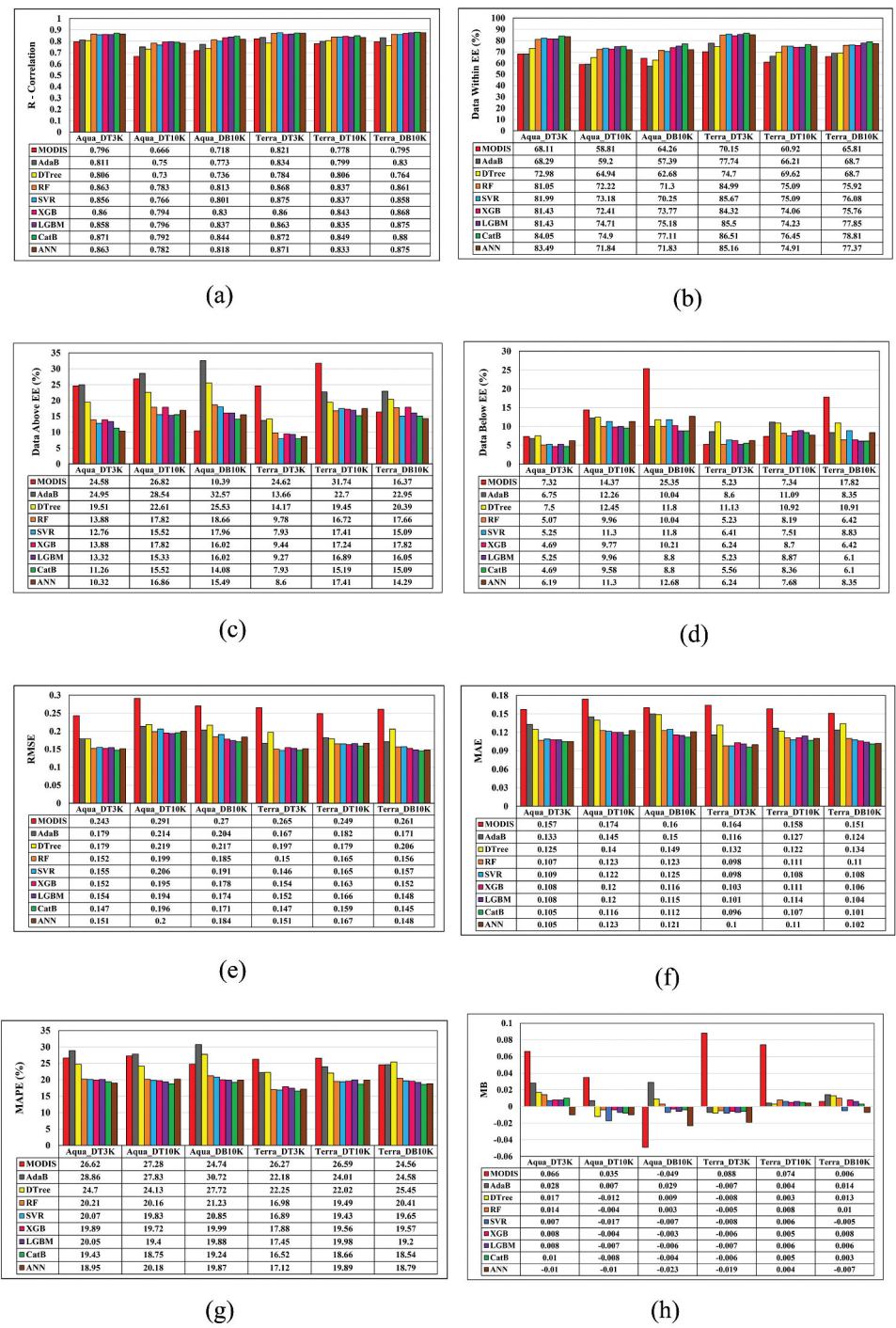


Figure 7. Performance metric-wise comparison of ML and ANN models for Aqua and Terra AOD products.

at 58.81%. Even in the best-case scenario for MODIS, the Terra $DT_{3\text{ km}}$ dataset, the percentage of data within EE reaches only 70.15%, still far below the performance of ML models. CatB consistently outperforms other ML models regarding the WEE across all MODIS aerosol retrieval products, with values ranging from 74.9% to 86.51%. Similar to the correlation analysis, DTree and AdaB show the lowest percentages of data within the EE boundary, highlighting their comparatively poorer performance. For the Aqua $DT_{3\text{ km}}$ dataset, ANN achieves the highest WEE of 83.49% next to CatB, while the remaining algorithms exhibit WEE values close to 81%. In the Aqua $DT_{10\text{ km}}$ dataset, LGBM achieves a high WEE of 74.71%, followed by SVR, XGB, RF and ANN. For Aqua $DB_{10\text{ km}}$, LGBM demonstrates strong performance with a WEE of 75.18%, followed by XGB, ANN, RF and SVR. In the Terra $DT_{3\text{ km}}$ dataset, SVR exhibits the highest WEE of 85.67%, followed by LGBM, ANN, RF and XGB. For the Terra $DT_{10\text{ km}}$ dataset, SVR, RF and ANN show WEE values close to 75%, indicating their reliability in this scenario. Similarly, for Terra $DB_{10\text{ km}}$, ANN and LGBM lead with high percentages of 77.37% and 77.85%, respectively, while the remaining models show WEE values near 76%. Among all the models, CatB, SVR, LGBM, and ANN demonstrate consistent performance, making them the most reliable choices for AOD prediction across MODIS aerosol retrieval products. In particular, CatB experienced the highest WEE (86.51) for Terra $DT_{3\text{ km}}$ among all the tested models and MODIS aerosol products.

Figure 7(c) indicates the percentage of data beyond the EE threshold, indicating prediction errors. MODIS AOD has the highest error rates, with Terra $DT_{10\text{ km}}$ being the worst (31.74%) and Aqua $DB_{10\text{ km}}$ the best (10.39%), but still lower than ML models. AdaB and DTree always have high errors, with AdaB being the highest at 32.57% and DTree at 25.53% for Aqua $DB_{10\text{ km}}$, as expected of their poorer accuracy. Except for Terra $DT_{10\text{ km}}$, RF and XGB exhibit the highest percentage of data over EE, followed by AdaB and DTree models. On the other hand, CatB, LGBM, SVR, and ANN have lower errors, with CatB and SVR having the best performance (7.93%) for Terra $DT_{3\text{ km}}$. ML models perform much better than MODIS, with CatB, SVR, LGBM and ANN being the most reliable for AOD prediction.

Figure 7(d) shows the percentage of data below the EE, reflecting the highly underestimating predictions for AOD values. MODIS records the highest rates of underestimation, most notably Aqua $DB_{10\text{ km}}$ (25.35%) and Terra $DB_{10\text{ km}}$ (17.82%), in support of its lower accuracy. Among the ML models, DTree and AdaB have consistently high errors, with DTree attaining 12.45% and AdaB reaching 12.26% for Aqua $DT_{10\text{ km}}$. Conversely, CatB and XGB have the lowest underestimation rates, at 4.69% for Aqua $DT_{3\text{ km}}$. ANN maintains excellent performance levels across all experiments with underestimation percentages of 6.19% (Aqua $DT_{3\text{ km}}$) and 12.68% (Aqua $DB_{10\text{ km}}$), demonstrating improved accuracy over MODIS and less accurate models such as DTree and AdaB. LGBM is also stable in its performance, with minimum values of 5.23% (Terra $DT_{3\text{ km}}$) and staying below 10%, further ensuring its dependability. SVR and RF remain competitive, registering one of the lowest underestimation values at 5.25% and 5.07% in Aqua $DT_{3\text{ km}}$, respectively. Generally, ML models minimize underestimation errors more than MODIS, producing the most accurate predictions.

Figure 7(e) highlights the RMSE values of these models, further measuring prediction accuracy. For Aqua $DT_{3\text{ km}}$, MODIS has the highest RMSE value of 0.243, indicating more significant prediction deviations, whereas CatB and ANN show smaller errors at 0.147 and

0.151, respectively. In Aqua $DT_{10\text{ km}}$, MODIS again records a high error of 0.291, while LGBM, XGB, and CatB demonstrate lower RMSE values of 0.194, 0.195, and 0.196, respectively. For Aqua $DB_{10\text{ km}}$, MODIS, as expected, maintains a high RMSE of 0.27, and CatB emerges as the most accurate with 0.171. Terra $DT_{3\text{ km}}$ presents a similar pattern with MODIS at 0.265, with CatB and SVR achieving a reduced error of 0.147 and 0.146, respectively. The Terra $DT_{10\text{ km}}$ dataset sees MODIS at 0.249, while XGB and CatB consistently achieve low RMSE values of 0.163 and 0.159. Finally, in the Terra $DB_{10\text{ km}}$ dataset, MODIS has an RMSE of 0.261, while CatB, LGBM, and ANN outperform others with values of 0.145, 0.148, and 0.148, respectively. Like previous metrics analysis, the DTree and AdaB produce higher RMSE values for all the MODIS datasets. Based on RMSE, models such as CatB, LGBM, XGB and ANN demonstrate the best prediction performance across all datasets. At the same time, MODIS consistently performs poorly in terms of accuracy and bias. Similar to R, Terra $DB_{10\text{ km}}$ has the minimum RMSE of 0.145 from CatB prediction.

Figure 7(f) compares the MAE of the different models, a key metric that reflects the average magnitude of prediction errors. Lower MAE values indicate more accurate predictions. For Aqua $DT_{3\text{ km}}$, MODIS has the highest MAE value of 0.157, suggesting a more significant prediction error, whereas CatB and ANN show minor errors of 0.105. In Aqua $DT_{10\text{ km}}$, MODIS again records a high error of 0.174, while CatB, LGBM, and XGB demonstrate lower MAE values of 0.116, 0.12 and 0.12, respectively. For Aqua $DB_{10\text{ km}}$, MODIS, as expected, maintains a high MAE of 0.16, and CatB emerges as the most accurate with 0.112. Terra $DT_{3\text{ km}}$ presents a similar pattern with MODIS at 0.164, with CatB achieving a reduced error of 0.096. The Terra $DT_{10\text{ km}}$ dataset sees MODIS at 0.158 of MAE, while SVR and CatB consistently achieve low MAE values of 0.108 and 0.107, respectively. Finally, in the Terra $DB_{10\text{ km}}$ dataset, MODIS has an MAE of 0.151, while CatB and ANN outperform others with values of 0.101 and 0.102, respectively. Like RMSE, the DTree and AdaB produce higher MAE values for all the MODIS datasets. The MAE analysis again proves that the CatB outperforms all the models, followed by ANN and LGBM. Compared to WEE, Terra $DT_{3\text{ km}}$ shows the lowest MAE value of 0.096 from CatB prediction compared to all the tested models among all the MODIS products.

Figure 7(g) presents the MAPE values for different models across various datasets. For the Aqua $DT_{3\text{ km}}$ dataset, MODIS has the highest MAPE value of 26.62%, while ANN has the lowest value at 18.95%, showing its superior accuracy, but AdaB has an increased MAPE (28.86%). Similarly, for Aqua $DT_{10\text{ km}}$, MODIS again has a high error of 27.28%, whereas CatB performs better with a lower value of 18.75%. Here, AdaB again has a higher MAPE of 27.83%. For Aqua $DB_{10\text{ km}}$, MODIS and AdaB record high values of 24.74% and 30.72%, while CatB and ANN show lower errors of 19.24% and 19.87%, respectively. Terra $DT_{3\text{ km}}$ demonstrates a similar trend, with MODIS at 26.27% and CatB at 16.52%, representing the best performance. Across Terra $DT_{10\text{ km}}$, MODIS consistently exhibits higher MAPE values, and models such as CatB, RF, and SVR achieve better accuracy with values around 18.66%, 19.49% and 19.43%, respectively. Similarly, for Terra $DB_{10\text{ km}}$, CatB and ANN show a lower MAPE of 18.54% and 18.79%, respectively. The CatB model shows the lowest MAPE value for all the datasets, followed by LGBM, ANN, and XGB. Overall, the Terra $DT_{3\text{ km}}$ shows a minimum MAPE of 16.52% among all the methods.

Figure 7(h) illustrates the MB values of the same models over the datasets. Positive and negative values indicate overestimation and underestimation tendencies, respectively. For Aqua $DT_{3\text{ km}}$, MODIS shows the highest bias of 0.066, significantly overestimating the

predictions, whereas SVR has a slight overestimation bias of 0.007. In the Aqua $DT_{10\text{ km}}$ dataset, MODIS exhibits a bias of 0.035, while models like XGB and RF lean towards minimal biases of -0.004 . For Aqua $DB_{10\text{ km}}$, MODIS underestimates with a bias of -0.049 , while RF and XGB maintain nearly unbiased predictions with values of 0.003 and -0.003 , respectively. On Terra $DT_{3\text{ km}}$, MODIS has the most substantial overestimation bias at 0.088, while RF displays a small negative MB value (-0.005). Across Terra $DT_{10\text{ km}}$, MODIS demonstrates biases of 0.074, while all the models show balanced biases ranging between 0.003 and 0.008, indicating a slight overestimation of the models. In contrast, AdaB and DTree show higher MB values than MODIS in Terra $DB_{10\text{ km}}$, while CatB denotes the slightly reduced MB value of 0.003. All the ensemble ML models showed reduced MB values for all the MODIS datasets. In particular, RF and XGB from Aqua $DB_{10\text{ km}}$, DTree from Terra $DT_{10\text{ km}}$, and CatB from Terra $DB_{10\text{ km}}$ show a minimum MB of 0.003.

It is concluded from the performance metric-wise analysis that the CatB model performs better in terms of all the parameters with higher correlation, a higher percentage of data within EE, lower error values, and a lower percentage of data above and below EE for both satellites and all the MODIS aerosol products. This research underscores the efficiency of advanced algorithms such as CatB for correcting bias, making them vital for enhancing satellite-based estimations of AOD. Likewise, ANN, LGBM, XGB, SVM, and RF also perform better in correcting MODIS-introduced bias. This highlights the ML models' capability to effectively correct biases in the original MODIS AOD product and demonstrates the advantage of using ML models to improve retrieval accuracy. It also considerably increases the validity of those predictions by retaining a higher percentage of data inside the predicted error range, and can manage complex, nonlinear relationships in the data. It is important to note that the model enhances existing AOD estimates rather than directly retrieving AOD values. The performance of DTree and AdaB models worsened for all the datasets, as indicated by higher error values, lower correlation, and lower percentage of data within EE. Ensemble-based methodologies and neural networks are notable tools for improving accuracy in satellite retrievals. Overall, Terra $DT_{3\text{ km}}$ and $DB_{10\text{ km}}$ show better performance in all metrics. In particular, CatB and Terra $DT_{3\text{ km}}$ are recommended for the Kanpur site's AOD prediction and analysis.

4.4. Comparison of bias corrected MODIS AOD using CatB against AERONET AOD

To provide a clear visual representation of how the ML model corrects the bias introduced by MODIS retrieval algorithms, we selected the best-performing model for each dataset and conducted a regression analysis comparing the model-predicted AOD and MODIS AOD against AERONET AOD. Based on prior analysis, CatB consistently demonstrated superior performance across all datasets. Figure 8 displays a comparative study of AOD at 550 nm, retrieved from MODIS observations and predicted using the CatB regression model, against ground-truth AOD measurements obtained from AERONET. The scatter plots are organized into six panels, representing different configurations of MODIS observations: (a) Aqua $DT_{3\text{ km}}$, (b) Aqua $DT_{10\text{ km}}$, (c) Aqua $DB_{10\text{ km}}$, (d) Terra $DT_{3\text{ km}}$, (e) Terra $DT_{10\text{ km}}$, and (f) Terra $DB_{10\text{ km}}$. Each panel compares the performance of MODIS-retrieved AOD (represented by red data points and trend lines) and CatB-predicted AOD (depicted by green data points and trend lines) against AERONET AOD, with a solid black 1:1 reference line indicating perfect agreement between actual and estimated values. The

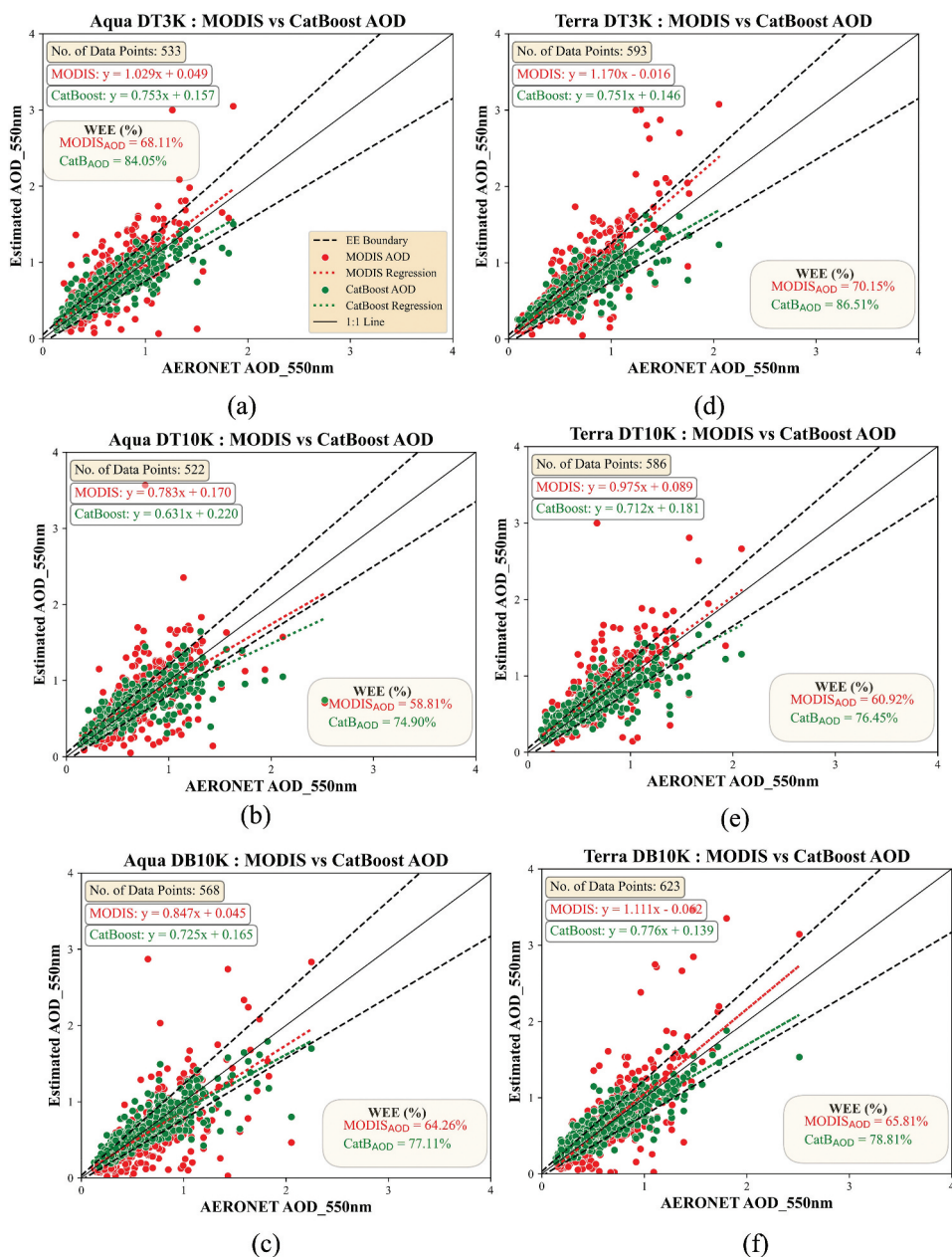


Figure 8. Comparison of CatB predicted AOD and MODIS retrieved AOD against AERONET AOD.

dashed black lines in all subplots are the EE envelope suggested by NASA's MODIS team, indicating retrieval accuracy. The legend depicted in Figure 8(a) applies to all subplots.

The regression equations and the percentage of data within the EE for MODIS (indicated by red) and CatB (green colour) predictions are provided in each subplot, and the total data points used. The percentage of data within the EE is a key metric for assessing the accuracy of MODIS and CatB AOD estimates compared to AERONET AOD. For Aqua

$DT_{3\text{ km}}$, 68.11% of MODIS-retrieved AOD values are within the EE, while CatB already improves this to 84.05%. Correspondingly, for Aqua $DT_{10\text{ km}}$, the percentage of EE for MODIS is 58.81%, while CatB is 74.90%, showing significant improvement. For Aqua $DB_{10\text{ km}}$, MODIS has 64.26% of data in EE, while CatB increases the same to 77.11%. For Terra $DT_{3\text{ km}}$, MODIS contains 70.15% of data in EE, and CatB increases the same to 86.51%. Similarly, Terra $DT_{10\text{ km}}$ indicates 60.92% of MODIS AOD in EE, and CatB enhances this to 76.45%. Lastly, in the case of Terra $DB_{10\text{ km}}$, MODIS is at 65.81% within EE, and CatB improves further to 78.81%. CatB predictions are more uniform throughout all retrievals, with a greater proportion of data points within the EE envelope, showing its ability to improve AOD retrieval accuracy beyond that of conventional MODIS techniques. Also, CatB predictions often show reduced scatter, indicating potentially lower error margins and outliers. Similarly, it demonstrates higher correlation and lower error metrics for all datasets than MODIS, validating its enhanced predictability. In conclusion, this study emphasizes the potential of ML-based models like CatB to complement traditional satellite retrieval methods for improved aerosol monitoring.

4.5. Computational efficiency analysis of ML and ANN models

Figure 9 compares the computational efficiency of various ML and ANN models for MODIS AOD products regarding computational time and memory usage. Figures 9(a–c) show the grid search, training, and testing times, while Figures 9(d–f) present the peak memory usage during grid search, training, and testing, respectively. ANN exhibits the highest runtime and memory usage across all MODIS AOD products, with peak memory exceeding 600 MB, highlighting its high computational cost despite strong predictive capability. In contrast, AdaB, SVR, and DTree show consistently low memory usage and minimal computational time across all phases, making them lightweight but less effective in predictive performance than advanced models. Ensemble methods such as RF, XGB and LGBM provide a balance between accuracy and computational cost but still require moderate resources. On the other hand, CatB clearly emerges as the most favourable model, delivering superior predictive performance with significantly lower computational cost than ANN and other ensemble models, and higher accuracy than all other models, making it the most practical and efficient choice for large-scale AOD bias correction. These accuracy and computational efficiency findings form a foundation for extending the framework to broader applications, leading to important directions for future investigation.

4.6. Future work

The present study focused on bias correction of MODIS AOD retrievals using ML and ANN models over a single site (Kanpur), which provided valuable insights into model performance and computational efficiency. However, aerosols vary highly across regions due to diverse emission sources, land-use patterns and meteorological conditions. Therefore, a natural extension of this work is to apply the proposed framework across a broader set of AERONET sites in India, particularly those in urban, rural, coastal, and desert regions, to assess the spatial generalizability of the models. This would enable evaluation of whether the superior performance of the models observed for Kanpur is site-specific or holds

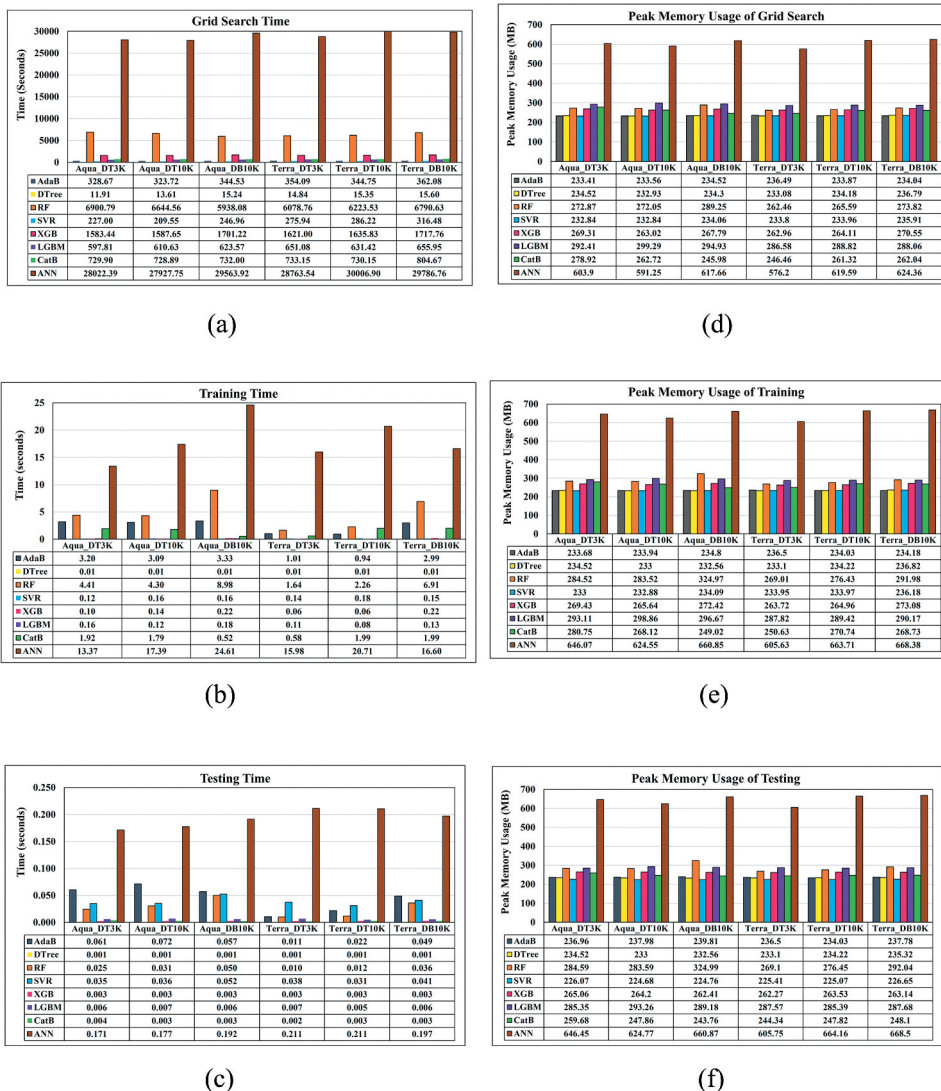


Figure 9. Computational time and memory usage performance of ML models: (a) grid search time, (b) training time, (c) testing time, (d) grid search memory, (e) training memory and (f) testing memory.

consistently across heterogeneous environments. At a global scale, incorporating multiple AERONET stations from varied climatological zones could help build a regionally adaptive or globally robust model for MODIS AOD bias correction.

Deep learning architectures such as CNNs, RNNs or hybrid DL – ML models could also capture complex spatiotemporal aerosol dynamics beyond what traditional ML models achieve. In addition to meteorological reanalysis variables, further improvement can be achieved by integrating supplementary MODIS parameters, such as solar and sensor zenith and azimuth angles, scattering angle, surface reflectance, TOA reflectance, and AOD at different wavelengths, which are known to influence AOD retrieval accuracy. Inclusion of other physical descriptors, such as cloud and land-cover information, may

also enhance model robustness under diverse aerosol regimes. Future studies should also explore multi-sensor fusion by combining MODIS with other satellite products like VIIRS, MISR and MAIAC, enabling better cross-platform consistency.

5. Conclusions

This paper proposes a comprehensive methodology to improve MODIS AOD retrievals by addressing biases and systematic errors using advanced ML and DL algorithms. The work is focused on data from the Kanpur AERONET station, spanning the entire MODIS measurement period for Terra (2001–2022) and Aqua (2002–2022) satellites. Using AERONET AOD as the reference, the proposed approach utilizes the latest MODIS Collection 6.1 data for bias correction. The methodology minimizes complexity by including only essential meteorological parameters and temporal variables while handling raw, unfiltered data effectively. The study emphasized the importance of feature selection by combining traditional feature importance rankings, permutation importance, and SHapley Additive Explanations (SHAP) values. These complementary methods offer a robust framework for identifying key predictors, leading to the selection of 13 critical features, including MODIS AOD, meteorological parameters (e.g. wind speed, wind direction, dew point temperature and air temperature at 2 metres, relative humidity, surface latent heat flux, boundary layer height, surface pressure, top and surface net thermal radiation), and temporal indicators (e.g. year and day of the year). This selection ensures that the models are accurate and efficient, avoiding unnecessary complexity. The selected data was applied to a variety of ML and DL algorithms such as CatBoost (CatB), LightGBM (LGBM), XGBoost (XGB), Random Forest (RF), Decision Tree (DTree), AdaBoost (AdaB), Support Vector Regression (SVR) and Artificial Neural Networks (ANN) to correct biases in multiple MODIS AOD products ($DT_{3\text{ km}}$, $DT_{10\text{ km}}$, $DB_{10\text{ km}}$) for both Terra and Aqua satellites.

Each algorithm was trained separately on the Terra and Aqua datasets, providing tailored corrections for different satellite measurements. Techniques such as 10-fold cross-validation, GridSearchCV for hyperparameter optimization, and advanced regularization methods like batch normalization, L2 regularization, and dropout were employed to ensure robust model performance. This careful implementation enhances model generalization, prevents overfitting, and provides stability in training. Performance evaluation of the models was carried out using a range of metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Correlation (R), Mean Absolute Percentage Error (MAPE), Mean Bias (MB), and the percentage of data within the error envelope (EE). The performance metric analysis concludes that the CatB model outperforms all other parameters, achieving higher correlations, a more significant percentage of data within the error envelope, and lower error values for both satellites and all MODIS aerosol products. Similarly, models like ANN, LGBM, XGB, SVM, and RF also demonstrate strong performance in correcting MODIS-induced biases.

In contrast, the DTree and AdaB models perform poorly across all datasets, as indicated by higher error values, lower correlations, and a reduced percentage of data within EE. CatB consistently outperformed others among the tested models, demonstrating its superior ability to handle non-linear relationships and high-dimensional data. For instance, the highest percentage improvements by the CatB model in terms of correlation,

WEE are up to 18.92% and as much as 27.36% for Aqua $DT_{10\text{ km}}$, respectively, showcasing its efficacy in addressing local biases and systematic errors. Similarly, the highest percentage reductions by the CatB model in terms of RMSE, MAE, and MAPE values are 44.53%, 41.46% and 37.11% respectively, for Terra $DT_{3\text{ km}}$. Also, the CatB model reduces the MB (overestimation) by a maximum of 122.86% for Aqua $DT_{10\text{ km}}$. Overall, CatB and Terra $DT_{3\text{ km}}$ are recommended for the Kanpur site's AOD prediction and analysis.

Acknowledgments

We thank NASA (National Aeronautics and Space Administration) and ECMWF (European Centre for Medium-Range Weather Forecasts) for providing useful AERONET, MODIS, and meteorological data for this research. We also thank the National Institute of Technology Puducherry, Karaikal, India, for providing research facilities.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

M. Anitha  <http://orcid.org/0000-0002-1748-0851>

Data availability statement

Data supporting the findings of this study are available from the corresponding author on reasonable request.

References

- Abd Jalal, K., A. Asmat, and N. Ahmad. 2015. "Aerosol Optical Depth (AOD) Retrieval Method Using MODIS." In *2015 International Conference on Space Science and Communication (ICONSPACE)*, Langkawi, 370–374. Malaysia: IEEE. <https://doi.org/10.1109/IconSpace.2015.7283802>.
- Akoshile, C. O., S. Shehu-Aladodo, M. Sani, J. O. Otu, and B. T. Ajibola. 2019. "Comparative Accuracy Assessment of Combined MODIS and NAAPS Aerosol Optical Depth with AERONET Data Over North Africa." *Atmospheric and Climate Sciences* 9 (3): 398–420. <https://doi.org/10.4236/acs.2019.93028>.
- Albayrak, A., J. Wei, M. Petrenko, C. Lynnes, and R. C. Levy. 2013. "Global Bias Adjustment for MODIS Aerosol Optical Thickness Using Neural Network." *Journal of Applied Remote Sensing* 7 (1): 073514–073514. <https://doi.org/10.1117/1.JRS.7.073514>.
- Anitha, M., and L. S. Kumar. 2020. "Ground Based Remote Sensing of Aerosols Using AERONET in Indian Region." In *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 72–77. Chennai, India: IEEE. <https://doi.org/10.1109/WiSPNET48689.2020.9198398>.
- Anitha, M., and L. S. Kumar. 2023a. "Tracking of NO₂ and SO₂ Trace Gases Emission from Thermal Power Plants in Tamil Nadu Using Sentinel 5P Tropomi Satellite with Observations from CPCB CAAQM Station." In *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IconSCEPT)*, 1–6. Karaikal, India: IEEE. <https://doi.org/10.1109/IconSCEPT57958.2023.10170014>.

- Anitha, M., and L. S. Kumar. 2023b. "Development of an IoT-Enabled Air Pollution Monitoring and Air Purifier System." *MAPAN – Journal of Metrology Society of India* 38 (3): 669–688. <https://doi.org/10.1007/s12647-023-00660-y>.
- Anitha, M., and L. S. Kumar. 2024. "An Analysis of Atmospheric Aerosol Characteristics Using Remote Sensing Data in the Indian Region." *Pure & Applied Geophysics* 181 (2): 625–654. <https://doi.org/10.1007/s00024-023-03415-7>.
- Annapurna, S. M., M. Anitha, and L. S. Kumar. 2024. "Composition and Source Based Aerosol Classification Using Machine Learning Algorithms." *Advances in Space Research* 73 (1): 474–497. <https://doi.org/10.1016/j.asr.2023.09.068>.
- Cao, Y., M. Wang, D. Rosenfeld, Y. Zhu, Y. Liang, Z. Liu, and H. Bai. 2021. "Strong Aerosol Effects on Cloud Amount Based on Long-Term Satellite Observations Over the East Coast of the United States." *Geophysical Research Letters* 48 (6): e2020GL091275. <https://doi.org/10.1029/2020GL091275>.
- Choi, W., H. Lee, and J. Park. 2021. "A First Approach to Aerosol Classification Using Space-Borne Measurement Data: Machine Learning-Based Algorithm and Evaluation." *Remote Sensing* 13 (4): 609. <https://doi.org/10.3390/rs13040609>.
- Choudhry, P., A. Misra, and S. N. Tripathi. 2012. "Study of MODIS Derived AOD at Three Different Locations in the Indo Gangetic Plain: Kanpur, Gandhi College and Nainital." *Annales Geophysicae* 30 (10): 1479–1493. <https://doi.org/10.5194/angeo-30-1479-2012>.
- Cuneo, L., A. G. Ulke, and B. Cerne. 2022. "Advances in the Characterization of Aerosol Optical Properties Using Long-Term Data from AERONET in Buenos Aires." *Atmospheric Pollution Research* 13 (3): 101360. <https://doi.org/10.1016/j.apr.2022.101360>.
- De Amorim, L. B., G. D. Cavalcanti, and R. M. Cruz. 2023. "The Choice of Scaling Technique Matters for Classification Performance." *Applied Soft Computing* 133:109924. <https://doi.org/10.1016/j.asoc.2022.109924>.
- Dubey, S. R., S. K. Singh, and B. B. Chaudhuri. 2022. "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark." *Neurocomputing* 503:92–108. <https://doi.org/10.1016/j.neucom.2022.06.111>.
- Dubovik, O., A. Smirnov, B. N. Holben, M. D. King, Y. J. Kaufman, T. F. Eck, and I. Slutsker. 2000. "Accuracy Assessments of Aerosol Optical Properties Retrieved from Aerosol Robotic Network (AERONET) Sun and Sky Radiance Measurements." *Journal of Geophysical Research Atmospheres* 105 (D8): 9791–9806. <https://doi.org/10.1029/2000JD900040>.
- ECMWF. 2024. "ERA5: How to Calculate Wind Speed and Wind Direction from U and V Components of the Wind?" Accessed November 20, 2024. <https://confluence.ecmwf.int/pages/viewpage.action?pageId=133262398>.
- Fan, Y., X. Sun, H. Huang, R. Ti, and X. Liu. 2021. "The Primary Aerosol Models and Distribution Characteristics Over China Based on the AERONET Data." *Journal of Quantitative Spectroscopy & Radiative Transfer* 275:107888. <https://doi.org/10.1016/j.jqsrt.2021.107888>.
- Gao, L., P. Kou, F. Gao, and X. Guan. 2010. "Adaboost Regression Algorithm Based on Classification-Type Loss." In *2010 8th World Congress on Intelligent Control and Automation*, 682–687. Jinan, China: IEEE. <https://doi.org/10.1109/WCICA.2010.5554029>.
- Gao, L., J. Li, L. Chen, L. Zhang, and A. K. Heidinger. 2016. "Retrieval and Validation of Atmospheric Aerosol Optical Depth from AVHRR Over China." *IEEE Transactions on Geoscience & Remote Sensing* 54 (11): 6280–6291. <https://doi.org/10.1109/TGRS.2016.2574756>.
- Giles, D. M., B. N. Holben, T. F. Eck, A. Sinyuk, A. Smirnov, I. Slutsker, R. R. Dickerson, A. M. Thompson, and J. S. Schafer. 2012. "An Analysis of AERONET Aerosol Absorption Properties and Classifications Representative of Aerosol Source Regions." *Journal of Geophysical Research Atmospheres* 117 (D17). <https://doi.org/10.1029/2012JD018127>.
- Global Monitoring Laboratory. 2019. "SURFRAD Aerosol Optical Depth." Accessed February 19, 2025. <https://www.esrl.noaa.gov/gmd/grad/surfrad/aod/>.
- Gong, C., J. Xin, S. Wang, Y. Wang, P. Wang, L. Wang, and P. Li. 2014. "The Aerosol Direct Radiative Forcing Over the Beijing Metropolitan Area From 2004 to 2011." *Journal of Aerosol Science* 69:62–70. <https://doi.org/10.1016/j.jaerosci.2013.12.007>.

- Hang, R., Q. Liu, G. Xia, and H. Song. 2018. "Correcting MODIS Aerosol Optical Depth Products Using a Ridge Regression Model." *International Journal of Remote Sensing* 39 (10): 3275–3286. <https://doi.org/10.1080/01431161.2018.1439597>.
- Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, et al. 2020. "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146 (730): 1999–2049. <https://doi.org/10.1002/qj.3803>.
- Hirtl, M., S. Mantovani, B. C. Krüger, G. Triebnig, C. Flandorfer, M. Bottoni, and M. Cavicchi. 2014. "Improvement of Air Quality Forecasts with Satellite and Ground Based Particulate Matter Observations." *Atmospheric Environment* 84:20–27. <https://doi.org/10.1016/j.atmosenv.2013.11.027>.
- Jabeur, S. B., C. Gharib, S. Mefteh-Wali, and W. B. Arfi. 2021. "Catboost Model and Artificial Intelligence Techniques for Corporate Failure Prediction." *Technological Forecasting and Social Change* 166:120658. <https://doi.org/10.1016/j.techfore.2021.120658>.
- Jiang, T., B. Chen, Z. Nie, Z. Ren, B. Xu, and S. Tang. 2021. "Estimation of Hourly Full-Coverage PM_{2.5} Concentrations at 1-km Resolution in China Using a Two-Stage Random Forest Model." *Atmospheric Research* 248:105146. <https://doi.org/10.1016/j.atmosres.2020.105146>.
- Just, A. C., M. M. De Carli, A. Shtein, M. Dorman, A. Lyapustin, and I. Kloog. 2018. "Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM_{2.5} in the Northeastern USA." *Remote Sensing* 10 (5): 803. <https://doi.org/10.3390/rs10050803>.
- Kim, M., S. H. Kim, W. V. Kim, Y. G. Lee, J. Kim, and M. C. Kafatos. 2021. "Assessment of Aerosol Optical Depth Under Background and Polluted Conditions Using AERONET and VIIRS Datasets." *Atmospheric Environment* 245:117994. <https://doi.org/10.1016/j.atmosenv.2020.117994>.
- Kumar, A., V. Pratap, S. Kumar, and A. K. Singh. 2022. "Atmospheric Aerosols Properties Over Indo-Gangetic Plain: A Trend Analysis Using Ground-Truth AERONET Data for the Year 2009–2017." *Advances in Space Research* 69 (7): 2659–2670. <https://doi.org/10.1016/j.asr.2021.12.052>.
- Lanzaco, B. L., L. E. Olcese, G. G. Palancar, and B. M. Toselli. 2016. "A Method to Improve MODIS AOD Values: Application to South America." *Aerosol & Air Quality Research* 16 (6): 1509–1522. <https://doi.org/10.4209/aaqr.2015.05.0375>.
- Lanzaco, B. L., L. E. Olcese, G. G. Palancar, and B. M. Toselli. 2017. "An Improved Aerosol Optical Depth Map Based on Machine-Learning and MODIS Data: Development and Application in South America." *Aerosol & Air Quality Research* 17 (6): 1623–1636. <https://doi.org/10.4209/aaqr.2016.11.0484>.
- Lary, D. J., L. A. Remer, D. MacNeill, B. Roscoe, and S. Paradise. 2009. "Machine Learning and Bias Correction of MODIS Aerosol Optical Depth." *IEEE Geoscience & Remote Sensing Letters* 6 (4): 694–698. <https://doi.org/10.1109/LGRS.2009.2023605>.
- Lemmouchi, F., J. Cuesta, M. Lachatre, J. Brajard, A. Coman, M. Beekmann, and C. Derognat. 2023. "Machine Learning-Based Improvement of Aerosol Optical Depth from CHIMERE Simulations Using MODIS Satellite Observations." *Remote Sensing* 15 (6): 1510. <https://doi.org/10.3390/rs15061510>.
- Li, C., J. Li, H. Xu, Z. Li, X. Xia, and H. Che. 2019. "Evaluating VIIRS EPS Aerosol Optical Depth in China: An Intercomparison Against Ground-Based Measurements and MODIS." *Journal of Quantitative Spectroscopy & Radiative Transfer* 224:368–377. <https://doi.org/10.1016/j.jqsrt.2018.12.002>.
- Lipponen, A., V. Kolehmainen, P. Kolmonen, A. Kukkurainen, T. Mielonen, N. Sabater, L. Sogacheva, T. H. Virtanen, and A. Arola. 2021. "Model-Enforced Post-Process Correction of Satellite Aerosol Retrievals." *Atmospheric Measurement Techniques* 14 (4): 2981–2992. <https://doi.org/10.5194/amt-14-2981-2021>.
- Liu, Y., T. Lin, J. Zhang, F. Wang, Y. Huang, X. Wu, H. Ye, G. Zhang, X. Cao, and G. de Leeuw. 2024. "Opposite Effects of Aerosols and Meteorological Parameters on Warm Clouds in Two Contrasting Regions Over Eastern China." *Atmospheric Chemistry & Physics* 24 (8): 4651–4673. <https://doi.org/10.5194/acp-24-4651-2024>.
- Loffe, S., and C. Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." arXiv preprint arXiv:1502.03167. <https://doi.org/10.48550/arXiv.1502.03167>.

- Malakar, N. K., D. J. Lary, A. Moore, D. Gencaga, B. Roscoe, A. Albayrak, and J. Wei. 2012. "Estimation and Bias Correction of Aerosol Abundance Using Data-Driven Machine Learning and Remote Sensing." In *2012 Conference on Intelligent Data Understanding*, 24–30. IEEE. <https://doi.org/10.1109/CIDU.2012.6382197>.
- Mangla, R., J. Indu, and S. S. Chakra. 2020. "Inter-Comparison of Multi-Satellites and AERONET AOD Over Indian Region." *Atmospheric Research* 240:104950. <https://doi.org/10.1016/j.atmosres.2020.104950>.
- Mhawish, A., T. Banerjee, D. M. Broday, A. Misra, and S. N. Tripathi. 2017. "Evaluation of MODIS Collection 6 Aerosol Retrieval Algorithms Over Indo-Gangetic Plain: Implications of Aerosols Types and Mass Loading." *Remote Sensing of Environment* 201:297–313. <https://doi.org/10.1016/j.rse.2017.09.016>.
- Misra, A., A. Jayaraman, and D. Ganguly. 2008. "Validation of MODIS Derived Aerosol Optical Depth Over Western India." *Journal of Geophysical Research Atmospheres* 113 (D4). <https://doi.org/10.1029/2007JD009075>.
- Misra, A., A. Jayaraman, and D. Ganguly. 2015. "Validation of Version 5.1 MODIS Aerosol Optical Depth (Deep Blue Algorithm and Dark Target Approach) Over a Semi-Arid Location in Western India." *Aerosol & Air Quality Research* 15 (1): 252–262. <https://doi.org/10.4209/aaqr.2014.01.0004>.
- Mohan, A. S., A. Manisekaran, and L. S. Kumar. 2021. "Aerosol Classification Using Machine Learning Algorithms." *Indian Journal of Radio & Space Physics* 50 (4): 217–223.
- More, S., P. Pradeep Kumar, P. Gupta, P. C. S. Devara, and G. R. Aher. 2013. "Comparison of Aerosol Products Retrieved from AERONET, MICROTOS and MODIS Over a Tropical Urban City, Pune, India." *Aerosol & Air Quality Research* 13 (1): 107–121. <https://doi.org/10.4209/aaqr.2012.04.0102>.
- Nirmalraj, S., A. S. M. Antony, P. Sridevionmalar, A. S. Oliver, K. J. Velmurugan, V. Elanagai, and G. Nagarajan. 2023. "Permutation Feature Importance-Based Fusion Techniques for Diabetes Prediction." *Soft Computing*: 1–12. <https://doi.org/10.1007/s00500-023-08041-y>.
- Olcese, L. E., G. G. Palancar, and B. M. Toselli. 2014. "Aerosol Optical Properties in Central Argentina." *Journal of Aerosol Science* 68:25–37. <https://doi.org/10.1016/j.jaerosci.2013.11.003>.
- Prasad, A. K., and R. P. Singh. 2007. "Comparison of MISR-MODIS Aerosol Optical Depth Over the Indo-Gangetic Basin During the Winter and Summer Seasons (2000–2005)." *Remote Sensing of Environment* 107 (1–2): 109–119. <https://doi.org/10.1016/j.rse.2006.09.026>.
- Rajendiran, N., and L. S. Kumar. 2023. "Pixel Level Feature Extraction and Machine Learning Classification for Water Body Extraction." *Arabian Journal for Science & Engineering* 48 (8): 9905–9928. <https://doi.org/10.1007/s13369-022-07389-x>.
- Rajendiran, N., S. Sebastian, and L. S. Kumar. 2024. "Cloud Segmentation, Validation of Weather Data, and Precipitation Prediction Using Machine Learning Algorithms." *Arabian Journal for Science & Engineering* 49 (9): 12259–12289. <https://doi.org/10.1007/s13369-023-08611-0>.
- Remer, L. A., R. G. Kleidman, R. C. Levy, Y. J. Kaufman, D. Tanré, S. Mattoo, J. V. Martins, et al. 2008. "Global Aerosol Climatology from the MODIS Satellite Sensors." *Journal of Geophysical Research Atmospheres* 113 (D14). <https://doi.org/10.1029/2007JD009661>.
- Sabetghadam, S., O. Alizadeh, M. Khoshsima, and A. Pierleoni. 2021. "Aerosol Properties, Trends and Classification of Key Types Over the Middle East from Satellite-Derived Atmospheric Optical Data." *Atmospheric Environment* 246:118100. <https://doi.org/10.1016/j.atmosenv.2020.118100>.
- Sangura, M. T., P. Althaf, J. W. Makokha, R. Boiyo, and K. R. Kumar. 2025. "Estimation and Model Performance of Aerosol Radiative Forcing from Radiative Transfer Models Using the AERONET Data Over Kenya, East Africa." *Advances in Space Research* 75 (1): 481–496. <https://doi.org/10.1016/j.asr.2024.09.061>.
- Sharma, V., S. Ghosh, M. Bilal, S. Dey, and S. Singh. 2021. "Performance of MODIS C6.1 Dark Target and Deep Blue Aerosol Products in Delhi National Capital Region, India: Application for Aerosol Studies." *Atmospheric Pollution Research* 12 (3): 65–74. <https://doi.org/10.1016/j.apr.2021.01.023>.
- Shi, S., T. Cheng, X. Gu, H. Guo, H. Chen, Y. Wang, and Y. Wu. 2018. "Multisensor Data Synergy of Terra-MODIS, Aqua-MODIS, and Suomi NPP-VIIRS for the Retrieval of Aerosol Optical Depth and Land Surface Reflectance Properties." *IEEE Transactions on Geoscience & Remote Sensing* 56 (11): 6306–6323. <https://doi.org/10.1109/TGRS.2018.2835508>.

- Tan, F., H. San Lim, K. Abdullah, T. L. Yoon, and B. Holben. 2015. "AERONET Data-Based Determination of Aerosol Types." *Atmospheric Pollution Research* 6 (4): 682–695. <https://doi.org/10.5094/APR.2015.077>.
- Tian, X., L. Gao, J. Li, L. Chen, J. Ren, and C. Li. 2021. "Retrieval of Atmospheric Aerosol Optical Depth from AVHRR over Land with Global Coverage Using Machine Learning Method." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–12. <https://doi.org/10.1109/TGRS.2021.3129853>.
- Tripathi, S. N., S. Dey, A. Chandel, S. Srivastava, R. P. Singh, and B. N. Holben. 2005. "Comparison of MODIS and AERONET Derived Aerosol Optical Depth Over the Ganga Basin, India." *Annales Geophysicae* 23 (4): 1093–1101. <https://doi.org/10.5194/angeo-23-1093-2005>.
- Vijayakumar, K., P. C. S. Devara, D. M. Giles, B. N. Holben, S. V. B. Rao, and C. K. Jayasankar. 2018. "Validation of Satellite and Model Aerosol Optical Depth and Precipitable Water Vapour Observations with AERONET Data Over Pune, India." *International Journal of Remote Sensing* 39 (21): 7643–7663. <https://doi.org/10.1080/01431161.2018.1476789>.
- Wang, H., Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar. 2024. "Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods." *Journal of Big Data* 11 (1): 44. <https://doi.org/10.1186/s40537-024-00905-w>.
- Wang, M., M. Fan, Z. Wang, L. Chen, L. Bai, Y. Chen, and M. Wang. 2023. "Machine Learning Based Bias Correction for MODIS Aerosol Optical Depth in Beijing." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-M-1-2023:395–402. <https://doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-395-2023>.
- Weber, S. A., J. A. Engel-Cox, R. M. Hoff, A. I. Prados, and H. Zhang. 2010. "An Improved Method for Estimating Surface Fine Particle Concentrations Using Seasonally Adjusted Satellite Aerosol Optical Depth." *Journal of the Air & Waste Management Association* 60 (5): 574–585. <https://doi.org/10.3155/1047-3289.60.5.574>.
- Yusuf, N., R. Said S, S. Tilmes, and E. Gbobaniyi. 2021. "Multi-Year Analysis of Aerosol Optical Properties at Various Timescales Using AERONET Data in Tropical West Africa." *Journal of Aerosol Science* 151:105625. <https://doi.org/10.1016/j.jaerosci.2020.105625>.
- Zaman, S. U., M. R. S. Pavel, K. S. Joy, F. Jeba, M. S. Islam, S. Paul, M. A. Bari, and A. Salam. 2021. "Spatial and Temporal Variation of Aerosol Optical Depths Over Six Major Cities in Bangladesh." *Atmospheric Research* 262:105803. <https://doi.org/10.1016/j.atmosres.2021.105803>.
- Zhang, J., and J. S. Reid. 2006. "MODIS Aerosol Product Analysis for Data Assimilation: Assessment of Over-Ocean Level 2 Aerosol Optical Thickness Retrievals." *Journal of Geophysical Research Atmospheres* 111 (D22). <https://doi.org/10.1029/2005JD006898>.