Air Quality, Atmosphere & Health (2022) 15:2185–2198
https://doi.org/10.1007/s11869-022-01245-5

Estimating daily $PM_{2.5}$ concentrations using an extreme gradient boosting model based on VIIRS aerosol products over southeastern Europe

Serdar Gündoğdu[1] · Gizem Tuna Tuygun[2] · Zhanqing Li[3] · Jing Wei[4] · Tolga Elbir[2]

Received: 4 January 2022 / Accepted: 26 August 2022 / Published online: 20 September 2022

Abstract

The performance of aerosol optical depth (AOD) products from the visible infrared imaging radiometer suite (VIIRS) instrument to estimate ground-level $PM_{2.5}$ concentrations has been determined at different locations; however, it is still limited over Europe. VIIRS dark target (DT) and deep blue (DB) AOD products at 6-km spatial resolution and independent variables from the MERRA-2 reanalysis were used for estimating daily $PM_{2.5}$ concentrations in southeastern Europe. An estimation model based on the Extreme Gradient Boosting (XGBoost) approach was developed and tested for DT and DB AODs. The estimations were compared with daily $PM_{2.5}$ observations from 122 air quality monitoring stations in five countries, including Bulgaria, Cyprus, Greece, Romania, and Turkey. The estimated $PM_{2.5}$ concentrations were consistent with ground measurements with the Pearson correlation coefficient (*R*) of 0.82 and 0.78, showing overall low estimation uncertainties with the root mean square error (RMSE) of 7.43 and 8.38 μg/m^3 and the mean absolute error (MAE) of 4.76 and 5.31 μg/m^3 for DT and DB AOD datasets, respectively. Independent model results were also discussed based on each country and season. The best estimation accuracy reached the *R* value of 0.83 with an average RMSE of 9.05 μg/m^3 and an MAE of 5.84 μg/m^3 in Turkey with DB AOD. In contrast, the model with DT AOD was highly accurate with the *R* value of 0.85, showing minor overall uncertainties (i.e., RMSE = 6.08 and 3.31 μg/m^3) over Greece. The highest accuracies were obtained in autumn and spring, while the lowest ones were available in winter and summer. This study provides a feasible machine learning approach to estimate $PM_{2.5}$ using VIIRS AOD products in southeastern Europe.

Highlights

- **XGBoost-based model was developed for estimating daily $PM_{2.5}$ concentrations in southeastern Europe**
- **The model performed better with VIIRS DT AOD in the region**
- **Highest estimation accuracies with DB and DT AODs were obtained over Turkey and Greece, respectively**
- **Autumn and spring had the highest accuracies, while the lowest accuracies were available in winter and summer**

Keywords $PM_{2.5}$ estimation · XGBoost · Aerosol optical depth · VIIRS · Southeastern Europe

✉ Gizem Tuna Tuygun
gizem.tuna@deu.edu.tr

[1] Computer Technology Program, Bergama Vocational High School, Dokuz Eylul University, Bergama-Izmir, Turkey

[2] Department of Environmental Engineering, Faculty of Engineering, Dokuz Eylul University, Buca-Izmir, Turkey

[3] Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA

[4] Department of Chemical and Biochemical Engineering, Center for Global and Regional Environmental Research, Iowa Technology Institute, University of Iowa, Iowa City, IA 52242, USA

Introduction

Atmospheric aerosols with an aerodynamic diameter of fewer than 2.5 μm ($PM_{2.5}$) are one of the most problematic air pollutants because of their adverse impacts on human health. These particles are of particular concern, as they can penetrate deeply into the lung, irritate, and corrode the alveolar wall and consequently impair lung function (Xing et al. 2016). Limited days with activity, premature mortality (Maji et al. 2018; Maciejewska 2020), increased number of people with heart or lung causes (Praznikar and Praznikar 2012; Yue et al. 2019), emergency room visits (Khan et al. 2019; Chen et al. 2020), respiratory symptoms with acute and chronic bronchitis, and asthma attacks (Xing et al. 2016; Liu et al. 2017; Jo et al. 2017) are associated with short-term exposure to $PM_{2.5}$. These aerosol particles are transported beyond geographic boundaries and contribute to the air quality on regional or global scales due to their smaller sizes (Kaneyasu et al. 2014; Wang et al. 2015).

Aerosol optical depth (AOD) provided by satellite remote sensing has proven to be a key predictor of ground-level particulate matter (PM) levels (Liu et al. 2007; Xue et al. 2019; Yazdi et al. 2020). The moderate-resolution imaging spectroradiometer (MODIS) AOD products of Terra and Aqua satellites have been extensively used to estimate PM concentrations due to their superior quality in the literature (Staggofia et al. 2017; Huang et al. 2018; Nabavi et al. 2019; Staggofia et al. 2019; Wei et al. 2019; Wei et al. 2020; Tuna Tuygun et al. 2021). The Terra satellite has been working longer than its designed operation time. Visible infrared imaging radiometer (VIIRS) started to operate on Suomi National Polar-orbiting Partnership (NPP) satellite as an extension of MODIS in 2011.

Different approaches have been widely used to obtain $PM_{2.5}$ concentrations establishing the VIIRS AOD–$PM_{2.5}$ relationship on different spatial scales (Wu et al. 2016; Yao et al. 2018, 2019; Gui et al. 2020; Wang et al. 2021; Wei et al. 2021). Deep learning techniques allow us to create models by leveraging multiple layers of artificial neural networks to extract advanced features from the input variables. With the flexible combination of remote sensing, land use, and meteorological inputs, a powerful learning capacity is constructed with a deep neural network. Nonlinear interactions among variables can be determined with advanced representations compared with many traditional models such as multiple linear regression (Zeydan and Wand 2019), generalized additive model, and support vector machines for regression (Liu et al. 2009). Recent studies show that several ensemble machine learning approaches based on random forest (Wei et al. 2019), extremely randomized trees (Wei et al. 2020; Wei et al. 2021), and eXtreme Gradient Boosting (XGBoost) seem to overcome the traditional methods for estimating $PM_{2.5}$ (Hu et al. 2017; Yan et al. 2020; Wei et al. 2021; Zhang et al. 2021).

Yao et al. (2018) indicated that a linear mixed effect model based on the VIIRS AOD could explain 76% of the $PM_{2.5}$ variations in the Beijing-Tianjin-Hebei region and provide better results than the model based on MODIS AOD (~71%). A virtual ground-based $PM_{2.5}$ observation network was constructed by Gui et al. (2020) with the XGBoost model across China. The model estimated daily $PM_{2.5}$ with high accuracy ($R^2 \sim 0.79$) across China. Wei et al. (2021) developed a space–time extremely randomized trees (STET) model with VIIRS AOD over China. The STET model produced highly consistent $PM_{2.5}$ estimations with ground-based measurements (CV-R^2 of 0.88) at the national scale. However, the VIIRS AOD products are mainly used in China to estimate $PM_{2.5}$ concentrations. There is no reported study for $PM_{2.5}$ estimation based on VIIRS AOD in the literature for the European region.

This study aimed to determine the performance of VIIRS AOD products to estimate ground-level $PM_{2.5}$ concentrations over southeastern Europe by the XGBoost model and test the model performance at spatial and temporal scales. The XGBoost is a decision tree-based ensemble machine learning algorithm widely used in data mining with high estimation success (Chen and Guestrin 2016). The primary attempt was to capture the spatiotemporal heterogeneity of model predictors. After testing the model performance, the spatial predictive capability of the model was determined. Then, country-based models were also proposed based on the inherent characteristics of ground-level $PM_{2.5}$ concentrations. The XGBoost model developed in this study is a first step to constructing a high-quality $PM_{2.5}$ dataset across southeastern European countries that are important for air pollution studies.

Datasets

Ground-based $PM_{2.5}$ measurements

Transport, industry, residential heating, production and distribution of energy, agriculture, waste landfill, waste incineration with heat recovery, and open burning of waste are the major sectors contributing to $PM_{2.5}$ emissions in Europe (EEA 2018). Natural sources also contribute to background PM concentrations with high PM levels due to desert dust transport and wildfires (EEA 2018).

The air quality monitoring stations were selected in the countries of southeastern Europe like Bulgaria, Cyprus, Greece, Romania, and Turkey (Fig. 1) and represent different environmental conditions and $PM_{2.5}$ levels. These stations are distributed unevenly, and most are clustered in the urban area. The availability of valid daily data was a crucial criterion in selecting the stations and the study period. As Turkey is the only exception that data were only obtained for 2018, the study period was selected from January 1 to December 31, 2018. Ground daily $PM_{2.5}$ observations were collected from the official website of the European Environment Agency (EEA)

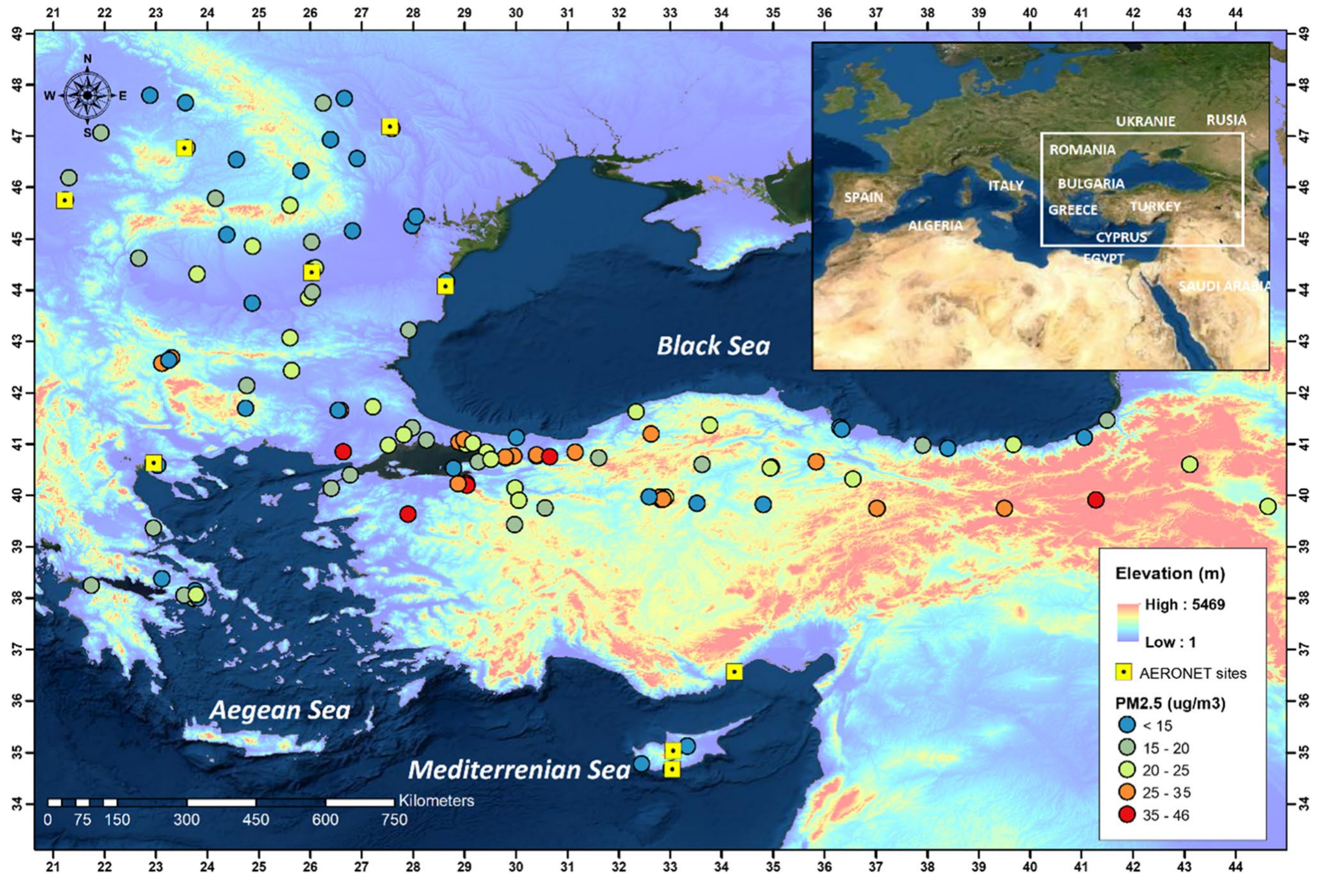

Fig. 1 Selected $PM_{2.5}$ monitoring stations (colored dots), mean $PM_{2.5}$ concentrations in 2018, and topography (background colored shading with GTOPO30 DEM) in the study area

(https://www.eea.europa.eu/data-and-maps/data/aqereporting-8). EEA provides European air quality information reported by all member countries and the cooperating and other reporting countries on this site. The tapered element oscillating microbalance method or the beta-attenuation method with appropriate calibration processes and quality controls are used to measure ground-level $PM_{2.5}$ concentrations in the region.

Based on the data availability, 122 $PM_{2.5}$ monitoring stations distributed in the study area were selected. Figure 1 presents the countries and stations included in this study and their location and annual average $PM_{2.5}$ concentrations over the region. The selection did not cover all available stations and areas across southeastern Europe. However, the stations showed wide distribution throughout the countries with different geographical and meteorological conditions. A list of the stations and a summary of descriptive statistics such as maximum, minimum, mean, and standard deviation are given in Table S1. According to the EU Directives on Ambient Air, the annual mean $PM_{2.5}$ concentration should be lower than 25 µg/m^3. In 2018, the overall country-based annual average $PM_{2.5}$ concentrations were not higher than the annual $PM_{2.5}$ limit value (25 µg/m^3). However, annual mean $PM_{2.5}$ concentrations above the limit value were observed in 24% ($n = 29$) of all the stations, mainly in Turkey ($n = 26$). The stricter standard of the World Health Organization (WHO) Air Quality Guidelines for annual mean $PM_{2.5}$ concentration (10 µg/m^3) was exceeded at 98% ($n = 120$) of all the stations.

Satellite AOD retrievals

The VIIRS is a new polar-orbiting sensor designed to substitute MODIS. Whereas the MODIS instrument provides high radiometric sensitivity in 36 spectral bands ranging in wavelength from 0.4 to 14.4 µm, VIIRS has 22 bands ranging from 0.412 to 12.01 µm. Furthermore, the day/night band is recently added to the VIIRS products to better determine night-time lights globally at a high spatial resolution. The VIIRS L2 products, including Deep Blue Aerosol L2 6-Min Swath 6 km (AERDB DB) and Dark Target Aerosol L2 6- Min Swath 6 km (AERDT DT), are available in public (Levy et al. 2015; Sayer et al. 2018; Sawyer et al. 2020). All VIIRS aerosol products are obtained at 6 × 6 km with the aggregation

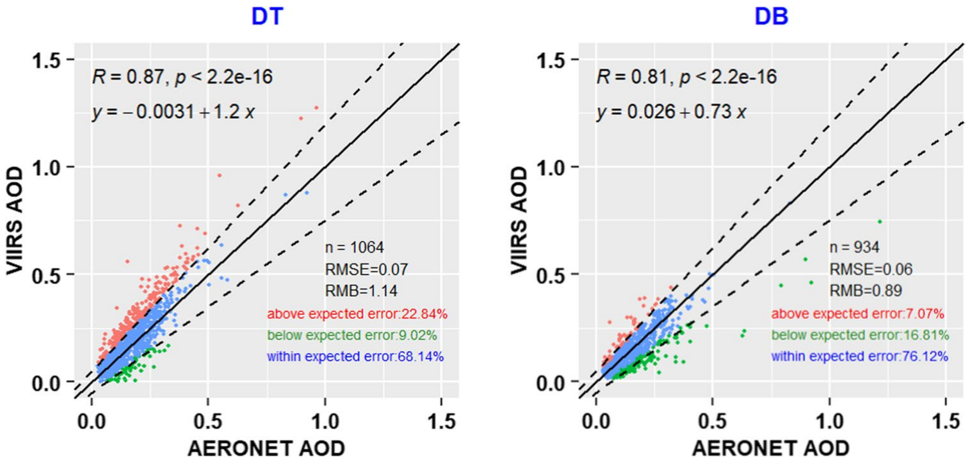

Fig. 2 Scatterplots of VIIRS DT and VIIRS DB AOD products vs. AERONET AOD in southeastern Europe

of 8 × 8 pixels of native VIIRS L1 products at ~ 750-m spatial resolution. The VIIRS AERDT DT and AERDB DB Level 2 AOD products for 2018, covering the entire study region, were used in this study. VIIRS DT AOD retrievals with a quality flag (QF) filter of QF > 1 for the ocean and QF = 3 for land and the high-quality VIIRS DB AOD retrievals (QF > 2) were used in this study. Only DT and DB AOD retrievals (550 nm) passing the quality assurance with the highest quality (i.e., quality assurance = best) were used as the primary independent variables to estimate $PM_{2.5}$ concentrations. In addition, Version 3 Level 2.0 AOD measurements from the AErosol RObotic NETwork (AERONET) at nine stations (Fig. 1) across the study region were obtained for the accuracy assessment of the VIIRS DT and DB AOD products in 2018.

The spatiotemporal approach, which was firstly proposed by Ichoku et al. (2002), was selected in this study to collocate data both spatially (averaged satellite retrievals in 5 × 5 grid boxes centered sunphotometer measurements of AOD) and temporally (within a time window of ± 30 min of satellite overpass time). The uncertainties for the matched data were determined to evaluate the VIIRS AOD quality. Figure 2 shows the relationship between the satellite-based and ground-based AODs in southeastern Europe. The expected error (EE) envelope representing approximately one standard deviation of the matchups (at least 67% of data points) is defined for each satellite algorithm. The EEs of the DT and DB algorithms are ($\pm 0.05 \pm 0.15 \times AOD_{AERONET}$) and ($\pm 0.03 \pm 0.20 \times AOD_{AERONET}$), respectively. The EE boundaries illustrated as dotted lines, AOD collocation pairs (*n*), Pearson correlation coefficient (*R*), root mean square error (RMSE), relative mean bias (RMB), and 1:1 line on a scatterplot are given in Fig. 2. The results indicated consistency between the VIIRS AOD retrievals and AERONET AODs with a high R of 0.87 and 0.81, a low RMSE of 0.07 and 0.06, and reasonable RMB of 1.14 and 0.89 for DT and DB products, respectively. Both DT and DB algorithms indicated a satisfactory performance of at least 67% of matchups falling within the EE range (68.14 and 76.12%, respectively). However, the VIIRS DT retrievals yield lower accuracy with 22.84% of collocated data above EE, showing overestimation. In summary, the quality of VIIRS AOD can provide a stable AOD-$PM_{2.5}$ relationship in the region.

Meteorological data

The Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2), was used to obtain surface mass concentrations of $PM_{2.5}$ components and meteorological data. The MERRA-2 is a global atmospheric reanalysis product produced by the NASA Global Modeling and Assimilation Office (GMAO) (Gelaro et al. 2017). Regularly gridded data with a homogeneous record of the global atmosphere and additional aspects of the climate system is provided by MERRA-2 (Gelaro et al. 2017; Randles et al. 2017).

Several parameters from the MERRA-2 reanalysis were obtained from the hourly averaged assimilated fields. Multiple meteorological and surface variables representing the relationship between the ground-level $PM_{2.5}$ and satellite-based AOD were collected in 2018. Hourly data of planetary boundary layer height (PBLH) (m), pressure (PS) (Pa), total cloud area fraction (CLDTOT) (unitless), total precipitation (PRECTOT) (kg/m^2/s), air temperature at 2 m (T2M) (K), relative humidity (RH) at a 1000-hPa surface pressure (1), evaporation (EVAP) (kg/m^2s), 10-m horizontal and vertical components of wind (U10M and V10M) (m/s), surface roughness (Z0M) (m), greenness fraction (GRN), and leaf area index (LAI) (1) were used as input variables in this study. Hourly surface mass concentrations (kg/m^3) of dust (DUSMASS25), black carbon (BCSMASS), organic carbon (OCSMASS), sulfate (SO4SMASS), and sea salt (SSSMASS25) were also used from MERRA-2. Daily average values were obtained from hourly observations, and these data had a spatial resolution of about 50 km. The spatial variables of longitude (Lon) and latitude (Lat) and the temporal variables of month and day representing the seasonal variability were also included in the model. In summary, the XGBoost model was enhanced with spatial and temporal information to determine more accurate estimation results in this study.

Methodology

Data processing and feature selection

Locations of $PM_{2.5}$ monitoring stations were used to match the AOD pixels for data integration. Spatiotemporal collocation was carried out for all variables to be matched in time and space by accounting for differences in the resolution and frequency of the variables. The MERRA-2 aerosol fields and meteorological data were matched to each VIIRS grid. Outputs from the 0.625° by 0.5° grid cell containing each $PM_{2.5}$ monitoring station were extracted. Within a spherical distance of 30 km (6 km × 5 grids) from each $PM_{2.5}$ monitoring site, valid AOD retrievals were extracted from the VIIRS swaths (in a 6-km resolution). A mean AOD value for each station and date was calculated if more than five valid AOD measurements were available within a search radius of 30 km. Briefly, all data from MERRA-2 and VIIRS AOD falling in one spatial grid located in the monitoring sites were used to match the datasets. The spatiotemporal regression matrix was prepared by integrating the geographical covariates, including AOD, meteorology, aerosol fields, and additional temporal and spatial covariates for the area in $PM_{2.5}$ estimation. Only the points including all the independent variables and daily $PM_{2.5}$ were evaluated in the sample selection process. Finally, 11,216 and 6401 matched data points with DT and DB AOD products were selected for modeling.

Since different methods are used to evaluate the importance of the explanatory variables in the prediction studies (Biecek and Burzykowski 2021; Luo et al. 2021), a variable importance analysis was performed to evaluate the contribution of each predictor in this study. This method is based on the F Score (feature score) measure, which simply summarizes the number of times each feature is used in the decision trees (Chen and Lin 2006; Zhang et al. 2020; Dai et al. 2022). First, all initial features were applied to the XGBoost model building; then, the importance provided an *F* score determining whether the feature was retained. Finally, the most significant twelve variables (AOD, PS, Z0M, BC, Day, Month, PBLH, GRN, LAI, T2M, LAT, LON) having the F-score over the threshold value were selected as the input parameters for the estimation model (Fig. 3).

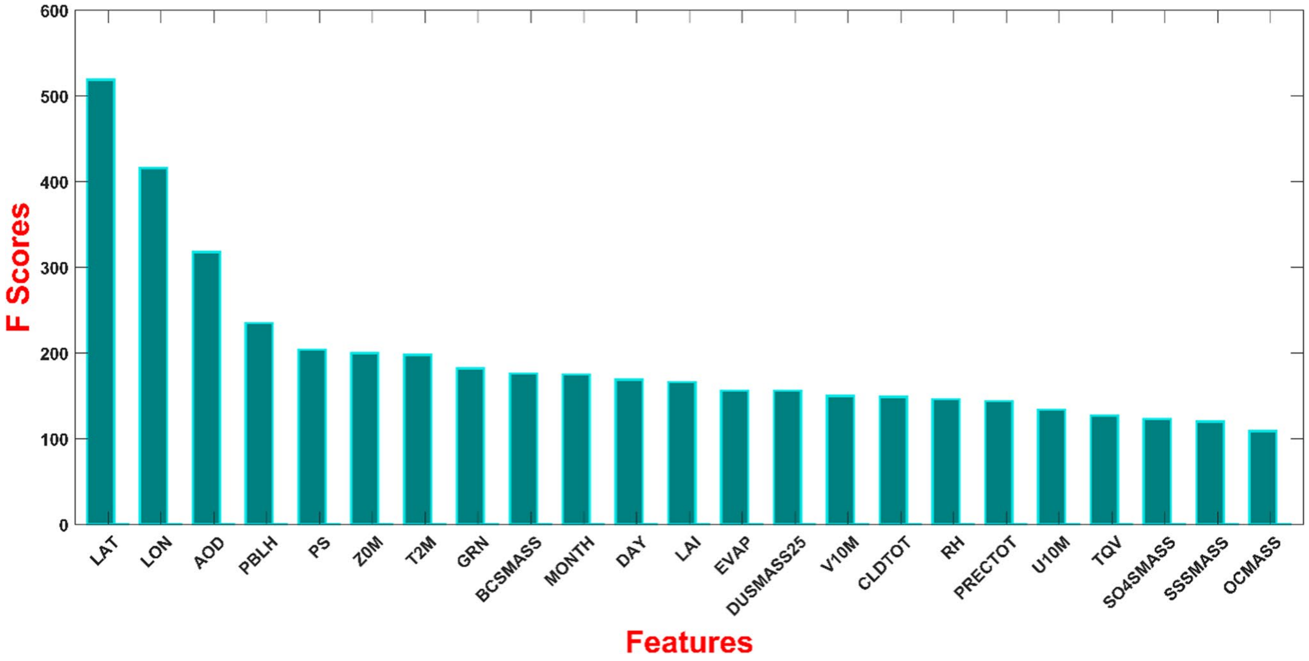

Fig. 3 Feature importance on $PM_{2.5}$ estimation of the XGBoost model

$PM_{2.5}$, as a dependent variable in the estimation model, is affected by various meteorological conditions (Czernecki et al. 2017; Jedruszkiewicz et al. 2017), topographical and land use variables (Choubin et al. 2020) commonly used in air quality studies. Although almost all variables widely used in the literature to estimate $PM_{2.5}$ were tried, as shown in Fig. 3, the *F* scores of eleven features (EVAP, RH, CLDTOT, U10M, V10M, DUSMASS25, SO4MASS, SSSMASS25, TQV, OCSMASS, and PRECTOT) indicated insignificant roles with weaker correlations than others. RH was expected to significantly affect AOD, which can mislead PM estimation but turns out to be relatively unimportant in the XGBoost model, as reported in Ghahremanloo et al. (2021). PRECTOT and RH cannot truly represent all atmospheric conditions since aerosol products are retrieved only at cloud-free pixels. This result is consistent with the study that estimated $PM_{2.5}$ in China (Wei et al. 2020). Moreover, the importance score in the XGBoost model only represents the importance of variables during the construction of the model and does not represent the physical contributions of the variables (Wei et al. 2020). Therefore, these variables were dropped from the modeling study.

The parameters of the spatial information had the highest score among all variables since the dataset includes $PM_{2.5}$ concentrations from 5 different countries. The *F* score of the AOD variable (318) was 35% higher than the closest follower variable (PBLH). AOD is the column integral of the atmospheric extinction coefficient and is the primary source for estimating surface PM concentrations. Zhang et al. (2018) also showed that AOD is the most significant factor in estimating $PM_{2.5}$ concentrations. As is well known from the literature, the ground-level $PM_{2.5}$ concentrations could not be estimated using AODs only. PBLH and PS are the other most crucial variables. The relationship between $PM_{2.5}$ concentrations and PBLH is driven by the vertical diffusion of pollutants. Higher pressure is usually associated with a low PBLH leading to high $PM_{2.5}$ concentrations (Li et al. 2018; Ghahremanloo et al. 2021). Temperature is also critical, and the formation of secondary aerosols could be promoted by low-temperature and high-humidity conditions. Previous studies have shown that the formation of secondary aerosols is the main reason for the growth of particulate matter (Chen et al. 2019). Meanwhile, temperature also affects the amount of coal burned in winter, which affects the anthropogenic emissions of $PM_{2.5}$.

Surface roughness, leaf area index, and greenness can affect spatial and temporal variability of air pollution. Barnes et al. (2014) showed that spatially and temporally varying surface roughness can significantly affect the ground-level air quality. Vegetation patterns influence the emission, diffusion, and absorption of PM (Feng et al. 2020). Land parameters such as topography, the normalized difference vegetation index (NDVI), and fractions of natural, agricultural, urban, or industrial areas are also significant factors (Stafoggia et al. 2019; Ghahremanloo et al. 2021). Due to their temporal limitations, surface roughness, leaf area index, and greenness were used as land variables instead of NDVI and land-use variables. Among the surface concentrations of OC, BC, DUST, and SS, only BC was the most significant component for $PM_{2.5}$ estimation, presumably indicating harmful particulate substances from combustion sources.

Considering the spatiotemporal heterogeneity is crucial to estimate accurate $PM_{2.5}$ estimations. Therefore, temporal features such as month and day of month were added to the final model. The final model was constructed with the top twelve input features, including AOD, PS, Z0M, BC, Day, Month, PBLH, GRN, LAI, T2M, LAT, and LON, expressed as follows:

$$PM_{2.5} = f(AOD, PS, Z0M, BC, Day, Month, PBLH, GRN, LAI, T2M, LAT, LON) \quad (1)$$

Since the spatial predictive power of the model was required, the training dataset was built by a randomly selected region, and the rest of the region was used as the test dataset. Finally, controlling the correlation between the observed and estimated $PM_{2.5}$ concentrations in held-out regions was done. Model evaluation was done with the testing data only. This method can provide different spatial points to create the training and testing samples from locations where the atmospheric and surface conditions can be noticeably different. The country-based model was also constructed to address the regional variation of the $PM_{2.5}$-AOD relationship and spatial heterogeneity of the model performance.

Application of XGBoost-based machine learning model for $PM_{2.5}$ estimation

An estimation model was developed using AOD data in two separate data sets (DT and DB) and auxiliary input variables determined by the *F* score to estimate daily mean full-coverage $PM_{2.5}$ concentrations in the PM monitoring sites. The model has consisted of matched mean daily values of all variables. The flow chart of the methodology used in this study is schematized in Fig. S1. In this methodology, AOD data with the other variables were taken as inputs, and $PM_{2.5}$ concentration was estimated as output. As mentioned in Sect. 3.1, the first stage is that the *F* score selected the significant independent variables. After that, these selected independent variables were used as the inputs of the machine learning model. Then, the Extreme Gradient Boosting (XGBoost) regression model was constructed to forecast $PM_{2.5}$ concentrations.

As its base learner, the XGBoost selects a decision tree. Adding new base learners decreases the estimation error, and the final estimated values are obtained by the mean of all base learners (Zheng and Wu 2019).

Assume that a dataset is DS = {($x_{j,}y_j$)} (j = 1,2,.....,*n*), and the model with *k* trees is learned (Song and Liu 2020). The result (p_j) of the model is as follows:

$$p_j = \sum_{k=1}^{K} f_k\left(x_j\right) f_k \in F \tag{2}$$

In Eq. (2), *f* is a regression tree, and *F* is the function of all decision trees.

$$F = \{f(x) = s_{q(x)}\} \tag{3}$$

where *q(x)* is the tree's leaf node of *x*th sample and *s* is the leaf score. When the *t*th step learning happens, the estimated result of x_j:

$$p_j^{\ t} = p_j^{\ t-1} + f_t(x_j) \tag{4}$$

In the process of regression, the objective function is:

$$Obj\left(f_t\right) = \sum_{j=1}^{n} L\left(y_j, p_j^{\ t}\right) + \Omega\left(f_t\right) \tag{5}$$

where *L* means loss function and Ω expresses the complexity of the model.

The details of θ(f_t) can be seen in Eq. (6):

$$\theta\left(f_t\right) = \gamma T_t + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^{\ 2} \tag{6}$$

where *T* and *w* indicate the number of a decision tree's leaf node and the score, respectively. Both *γ* and λ express the penalty factor. Finally, the objective function is:

$$Obj\left(f_t\right)^t = \sum_{i=1}^{n} L\left(y_i, p_j^{\ t}\right) + \Omega\left(f_t\right) = \sum_{i=1}^{n} L(y_i, p_j^{\ t-1} + f_t\,(x_j)) + \Omega\left(f_t\right) \tag{7}$$

XGBoost applies the greedy algorithm to build the decision tree based on the objective function. A complete XGBoost model is established by building decision trees constantly (Zheng and Wu 2019).

The procedure for working with the machine learning model includes three stages: (a) training, (b) testing of the model, and (c) validation of the estimations. In this study, the datasets were randomly divided into training and testing sets during the training process and randomly chosen 80% of the data for training and the remaining 20% for model testing. Splitting the dataset into other ratios such as 75–25%, 70–30%, 65–35%, and 60–40% was also tried in the study, and very close results were obtained in these trials. Therefore, the ratio of 80–20% was preferred in this study since it gave slightly better results than the other ratios and has been frequently used in similar PM prediction studies (Park et al. 2019; Zhao et al. 2019; Lv et al. 2021; Tuna Tuygun et al. 2021; Carreño et al. 2022). Phyton 3.7 and R 3.5.2 were used for all calculations and plot generations. The performance of the XGBoost model in estimating $PM_{2.5}$ levels was determined by several statistical indicators such as R, RMSE, and MAE, as expressed in Tuna Tuygun et al. (2021).

Results and discussions

Descriptive statistics

This section presents the descriptive statistics of the parameters used in this study to estimate daily ground-level $PM_{2.5}$ concentrations. As an overview of the variables used in this study, the histograms of all variables, including $PM_{2.5}$, are illustrated in Fig. S2 and S3 for the DT and DB datasets. Figures show that $PM_{2.5}$ concentrations ranged from 0.03 to 19.2 μg/m^3 and 0.49 to 20.4 μg/m^3, respectively. The annual mean $PM_{2.5}$ concentrations were 19.2 and 20.4 μg/m^3 for the datasets including DT AOD and DB AOD, respectively. The mean, median, and standard deviation of $PM_{2.5}$ concentrations for DT and DB datasets were similar (Fig. S2 and S3). The annual mean $PM_{2.5}$ concentrations over the region were much lower than that of other regions, such as China, with an annual mean of 51 μg/m^3 (Wei et al. 2019). The number of data with low $PM_{2.5}$ concentrations was significantly higher than those with high $PM_{2.5}$ concentrations. The annual mean measured $PM_{2.5}$ concentrations over the region differed for individual countries and seasons due to the combination of the different $PM_{2.5}$ emission sources (traffic, combustion, industry, traffic, mineral dust, sea salt, etc.) and the other factors such as meteorology, geography, or economy (Houthuijs et al. 2001; Diapouli et al. 2017; Nastase et al. 2018; Almeida et al. 2020). Although the mean observed $PM_{2.5}$ concentrations for the entire study period is approximately 20 μg/m^3, it differs for individual countries and seasons. Ground-level $PM_{2.5}$ concentrations showed a decreasing trend from winter to summer. The annual means of observed $PM_{2.5}$ concentrations in Turkey were higher than those in other countries, while lower annual mean $PM_{2.5}$ concentrations were observed in Cyprus (Fig. S4).

The frequency distributions of $PM_{2.5}$ and AOD were similar, indicating that the two parameters are related. The VIIRS-derived AOD data had a mean value of 0.156 and 0.197 and a standard deviation (SD) of 0.116 and 0.145 for DB and DT, respectively. Generally, AOD values between 0.01 and 0.4 are defined as extremely clean and hazy atmospheric conditions (NOAA 2020). Results indicated that the mean AOD values for the study period over the region are not significantly low or high and represent the relatively clean conditions. Moreover, AOD ranges from 0 to 1.954 and 0.013 to 1.664 for DT and DB datasets, respectively. The DT product obtained higher AOD values over the region. AOD also showed significant seasonal and country-level variability. VIIRS AOD values showed an increasing trend from winter to summer. Higher mean AOD values were observed in Cyprus, while low values were observed in Bulgaria (Fig. S4). Precipitation lowered AOD in winter in the region (Ozdemir et al. 2020). However, the largest mean $PM_{2.5}$ concentrations were observed in winter, and the lowest was in summer. Adverse meteorological conditions like lack of precipitation, low boundary layer height, and high aerosol emission explained the peak $PM_{2.5}$ concentration observed in winter (Barmpadimos et al. 2012; Adaes and Pires 2019; Tuna Tuygun and Elbir 2020). Both autumn and spring had similar $PM_{2.5}$ concentrations. The partial contribution of intense Saharan dust advection and deposition episodes that affected this area in spring is one of the reasons for higher $PM_{2.5}$ and AOD levels (Kaskaoutis et al. 2019; Achilleos et al. 2020; Ozdemir et al. 2020).

Model performance over southeastern Europe

Figure 4 shows the scatterplots depicting the relationship between measured and estimated $PM_{2.5}$ concentrations based on VIIRS DT and DB AOD products for training and test datasets. The estimated $PM_{2.5}$ concentrations are consistent with ground measurements with R of 0.82 and 0.78, showing overall low estimation uncertainties with the RMSE of 7.43 and 8.38 µg/m^3 and the MAE of 4.76 and 5.31 µg/m^3 for DT and DB, respectively. Similarly, the regression lines also have moderate slopes of 0.67–0.62 and small *y*-intercepts of 6.4 and 7.8 µg/m^3. Furthermore, the test results decreased smaller in most evaluation indexes with DB AOD, further demonstrating the robustness of

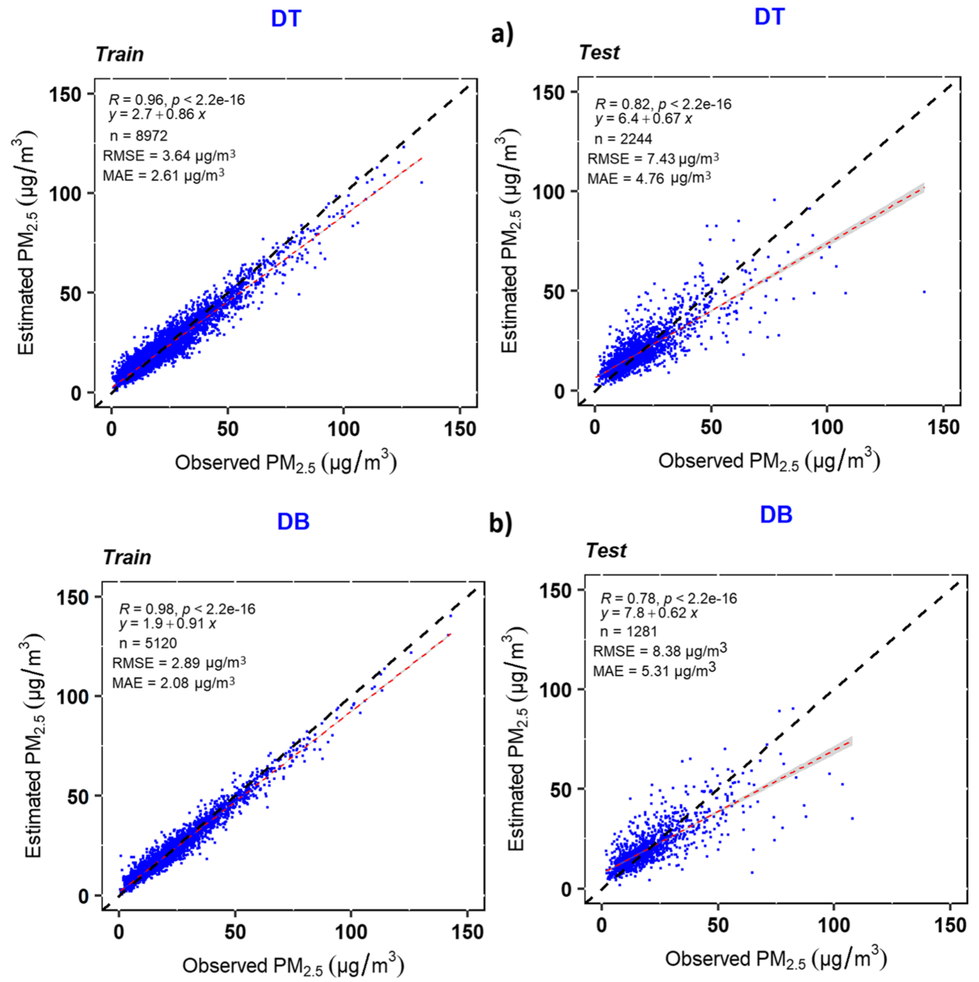

Fig. 4 XGBoost model results for train and test datasets with VIIRS DT and DB AOD products

the model. It is possibly due to different $PM_{2.5}$ and AOD loadings and the amount of data in the datasets. An R-value higher than 0.70 is considered significant in the literature (Ahmad et al. 2019) and represents a strong positive linear relationship between the observed and estimated values. Therefore, the XGBoost model can handle random variations in $PM_{2.5}$ concentrations and is considered acceptable in southeastern Europe. As shown in Fig. 4, the differences in R/RMSE/MAE between the train and test datasets indicate a slight over-fitting in the model. Compared to the previous related works in Europe, the estimation accuracy of the method developed in this study is satisfactory (Stafoggia et al. 2019; Zeydan and Wang 2019). Considering calibration of the $PM_{2.5}$-AOD relationship, the developed model in this study surpassed the previously published study (Zeydan and Wang 2019) to estimate the $PM_{2.5}$ concentrations in Turkey. Country-based validation results showed that model prediction ability over Turkey could achieve higher values ($R > 0.80$) with VIIRS DB AOD. Unlike this study, the most recent research used the AOD from the MAIAC algorithm, which showed much more accurate validation results than the DT and DB algorithms, notably more minor estimation uncertainties, mainly over complex urban areas (Wei et al. 2019). MAIAC AOD also has a higher spatial resolution (1 km) than commonly used MODIS DT (3–10 km) and DB (10 km) products. Stafoggia et al. (2019) explained the $PM_{2.5}$ variability over Italy, with mean CV- R^2 0.86 for $PM_{2.5}$ with gap-filled MAIAC data. Another study developed an ensemble model over Italy found R^2 as 0.81 for $PM_{2.5}$ with MAIAC AOD. Beloconi et al. (2018) compared geostatistical, geographically weighted, and land-use regression formulations over 46 European countries with MAIAC AOD. The geostatistical regression model was found as the most effective model over the region, with an R^2 of 0.78.

Figure 5 shows (a) time-series graphs of the observed and estimated $PM_{2.5}$ concentrations, and (b) the AOD-based $PM_{2.5}$ estimations bias by five quantile classes. Quantile class calculated in this study represents the 0.25, 0.50, 0.75,0.95, and 0.98, respectively. However, the quantiles may be more appropriate to represent uncertainty in this study, as Yan et al. (2021) preferred. Low-level $PM_{2.5}$ values mainly tended to overestimate with the XGBoost model. Di et al. (2016) also showed the negative impacts of low $PM_{2.5}$ concentrations on the model performance. However, it is obvious in Fig. 5 that $PM_{2.5}$ concentrations were both underestimated and overestimated by the XGBoost model. Positive and negative errors in the $PM_{2.5}$ estimations indicate that the XGBoost neither highly overestimated nor underestimated the $PM_{2.5}$ concentrations. High underestimations were observed for high $PM_{2.5}$ concentrations.

High bias and low model performance were found for upper percentiles (95th and 98th) of $PM_{2.5}$ (representing here as Quantile 4, Q4 and Quantile 5, Q5). Zhao et al. (2020) also found that the 95th and 98th percentiles of the $PM_{2.5}$ data yielded similar results. Many models currently show poor capabilities in estimating high values. These results indicate that $PM_{2.5}$ estimations still have significant uncertainty for areas and times of heavy pollution. Since clouds restrict the surface and make detecting aerosols on the ground impossible, satellites do not provide AOD information under cloudy conditions. However, the ground-level PM concentrations are measured continuously regardless of sky conditions (Park et al. 2020). Therefore, satellite-estimated daily PM concentrations with data gaps tend to underestimate

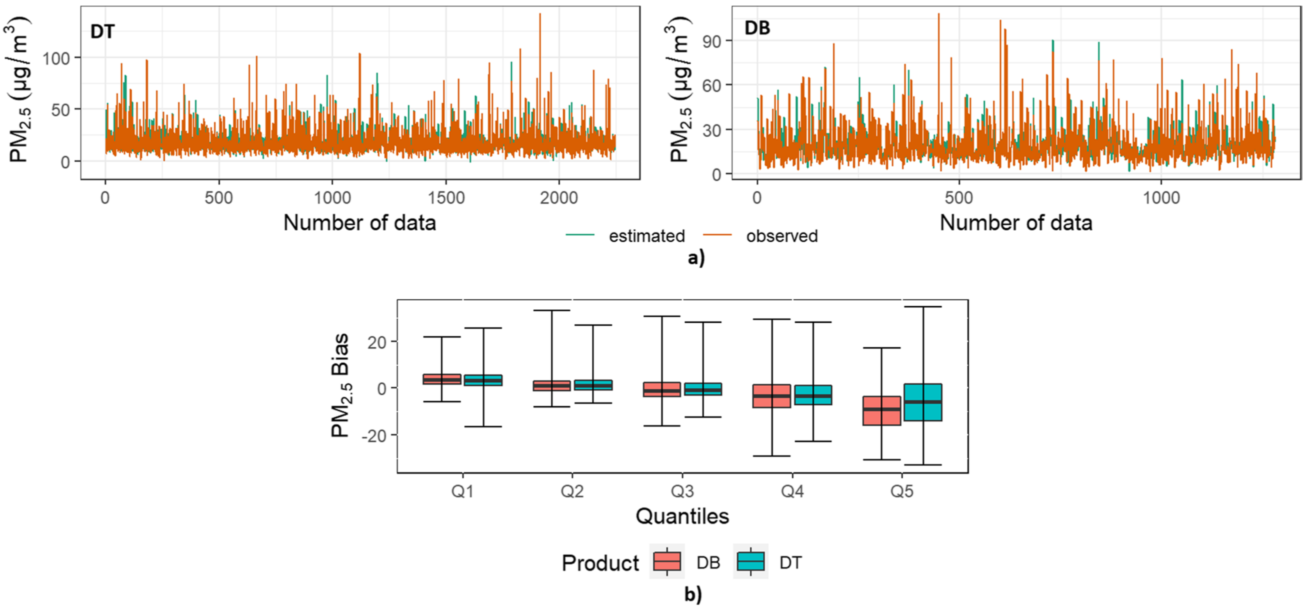

Fig. 5 Variation of the observed and predicted daily $PM_{2.5}$ concentrations by the DT and DB AODs (**a**) and $PM_{2.5}$ bias for quantile class (**b**)

Table 1 The spatial predictive power of the XGBoost model on each country

Country		DB				DT			
		N	*R*	RMSE	MAE	*N*	*R*	RMSE	MAE
Bulgaria	Train	5696	0.48	11.68	7.16	10,111	0.45	10.81	6.81
	Test	705				1105			
	All	6401				11,216			
Cyprus	Train	6038	0.27	6.58	5.10	10,600	0.44	5.51	4.21
	Test	363				616			
	All	6401				11,216			
Greece	Train	5912	0.39	10.52	6.73	9986	0.37	9.10	5.67
	Test	489				1230			
	All	6401				11,216			
Turkey	Train	3400	0.38	16.01	10.32	5502	0.45	14.09	8.94
	Test	3001				5654			
	All	6401				11,156			
Romania	Train	4558	0.43	9.14	6.46	8605	0.43	8.70	6.09
	Test	1843				2611			
	All	6401				11,216			

the concentrations compared with ground-level measurements (Xiao et al. 2017; Stafoggia et al. 2019; Wei et al. 2019). A few studies have tried to solve this challenge by combining satellite-derived and model-simulated AOD to fill gaps (Xiao et al. 2017; Stafoggia et al. 2019; Jiang et al. 2021; Tuna Tuygun et al. 2021). Xiao et al. (2017) combined MAIAC with chemical transport model (CTM) simulations in Yangtze River Delta, China, and Stafoggia et al. (2019) filled missing MAIAC AOD with Copernicus Atmosphere Monitoring Service (CAMS) in Italy.

All geographical regions were respectively excluded and used as test data to determine the spatial predictive power of the model. Table 1 shows which regions were used for training and estimated during each model run. The spatial predictive power of the XGBoost model has poor accuracies (R ~0.27–0.48) since AOD and $PM_{2.5}$ concentrations are not noticeably spatially changed over the different countries. The daily $PM_{2.5}$ estimates were not highly consistent with ground measurements, with high uncertainties and low R in all countries. By contrast, Cyprus showed poor overall accuracy based on DB AOD with low R values due to the sparse site distributions and low $PM_{2.5}$-polluted conditions. The XGBoost model generally showed a stable performance with AOD products in Bulgaria, Greece, and Romania. However, all regions obtained lower RMSE values with DT AOD.

Test results of daily estimated $PM_{2.5}$ concentrations against the ground measurements in each country are shown in Fig. 6. The results indicated that the $PM_{2.5}$ estimates were moderately correlated with the ground-level concentrations, ranging R values from 0.58 to 0.85 across the study region. The best estimation accuracy of the XGBoost model with

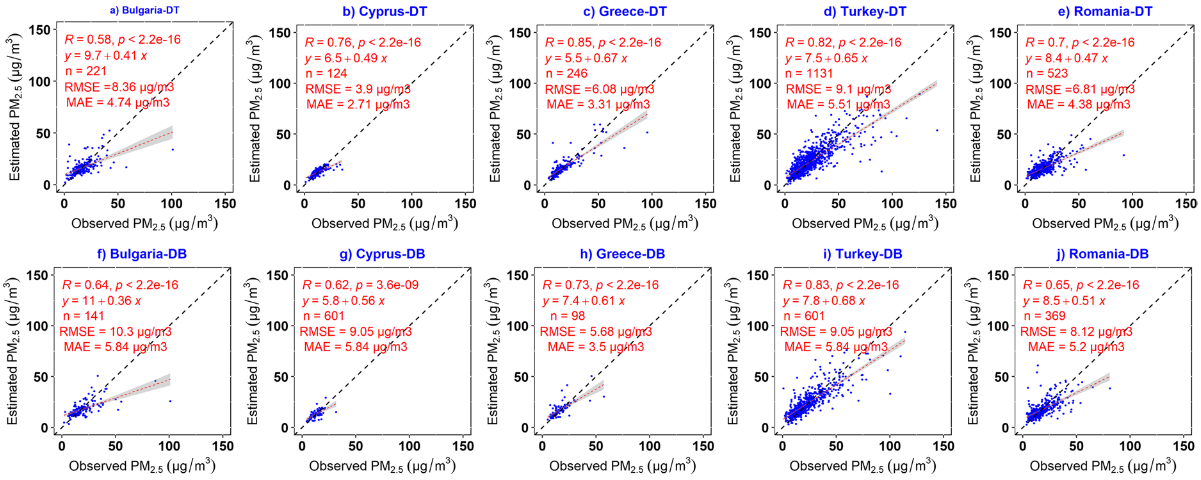

Fig. 6 Country-based model performances with the XGBoost model based on DT AOD (**a**–**e**) and DB AOD (**f**–**j**)

Table 2 Seasonal results of the XGBoost model

Product	Season	R		RMSE		MAE	
		Train	Test	Train	Test	Train	Test
DT AOD	Spring	0.99	0.81	1.59	6.61	1.14	4.57
	Summer	0.97	0.77	1.56	3.92	1.15	2.81
	Autumn	0.99	0.81	2.39	7.83	1.72	5.17
	Winter	0.99	0.69	3.20	15.54	2.32	10.69
DB AOD	Spring	0.98	0.75	2.64	7.65	1.94	5.06
	Summer	0.96	0.69	1.80	4.90	1.35	3.23
	Autumn	0.99	0.79	2.24	8.41	1.55	5.87
	Winter	1.00	0.66	0.67	16.95	0.46	12.99

DB product reached 0.83 with an average RMSE of 9.05 μg/m^3 and an MAE of 5.84 μg/m^3 in Turkey. However, the XGBoost model with DT AOD was highly accurate with R values of 0.85, showing minor overall uncertainties (i.e., RMSE = 6.08 and 3.31 μg/m^3) over Greece. These results highlighted the accuracy differences in various countries in the region. Differences and difficulties in model performances over the region are considered to be originated from the climate conditions (e.g., precipitation and high humidity) and the variances in a number of the data samples.

The training and estimation performance (test) results in each season are presented in Table 2. In terms of test results, autumn and spring had the highest accuracies. While the performance of the model slightly varied by season, the XGBoost model sufficiently estimated $PM_{2.5}$ concentrations at the seasonal scale. Conversely, winter showed the lowest accuracies ($R = 0.69/0.66$, RMSE = 15.54/16.95 μg/m^3, MAE = 11.64/12.99 μg/m^3) with DT/DB AODs. High cloud fraction conditions can cause negative impacts on the performance of the model in winter (Guo et al. 2017). Winter demonstrated similar performance with both products due to the higher concentrations in winter. The developed model in this study also underestimated the higher concentrations over the region (Fig. 5).

Conclusion

The first attempt in the literature was to develop the XGBoost model to estimate the spatial and temporal variations of $PM_{2.5}$ levels in five southeastern European countries by incorporating VIIRS AOD instead of commonly used MODIS AOD products. The XGBoost model performance was satisfactory, providing daily air quality assessment on a regional scale. The model was tested with ground-based measurements and comprehensively evaluated the DT and DB AOD-based $PM_{2.5}$ concentrations. Both DT and DB retrievals showed similar accuracy, suggesting that the XGBoost model performed stably. However, DT AOD represented the region better.

Satellites are a reliable data source for many research areas because of their high spatial coverage and long-term data products. However, missing AOD values restrict their usage in PM estimation. The accurate estimation of $PM_{2.5}$ distributions is difficult since many factors affect the ground-level $PM_{2.5}$ concentrations due to the temporal and spatial variability of pollution sources and atmospheric conditions. In this study, the most influential factors affecting $PM_{2.5}$ were considered. F score of AOD obtained by feature importance indicates that AOD was an essential variable for estimating ground $PM_{2.5}$ concentrations over the region after the two spatial variables. Moreover, a comparison between the model results with the DT and DB AOD datasets showed that the estimation accuracy did not vary significantly for different AOD datasets.

The XGBoost models incorporating satellite-derived data to estimate $PM_{2.5}$ concentrations have been developed in China in recent years (Meng et al. 2016; Zhang et al. 2018; Chen et al. 2019; Wei et al. 2019, 2020; Wei et al. 2021) while the application of AOD in estimating ground-level PM concentrations is limited over Europe. Only a few studies estimating the $PM_{2.5}$ concentrations by different machine learning and linear mixed effect models in Italy are available (Stafoggia et al. 2017, 2019; Shtein et al. 2019). They filled the missing AOD data with CAMS AOD data. The AOD data were not filled in this study, and it should be noted that these missing AOD data can still affect the estimation performance. Therefore, more accurate determination approaches need to be investigated to improve the spatio-temporal accuracy of $PM_{2.5}$ estimates. On the other hand, other studies estimated PM in Europe without using satellite-based AOD data (Czernecki et al. 2021). Czernecki et al. (2021) tested different ML models (AIC-based stepwise regression, two tree-based algorithms, and neural networks) for forecasting PM_{10} and $PM_{2.5}$ concentrations in 11 stations in Poland. They found that all methods obtained high accuracies, but the XGBoost performed the best, followed by random forests and neural networks, and stepwise regression performed the worst.

Previous studies employing statistical models were conducted at the national and regional scales, and a few compared the variability of $PM_{2.5}$ estimation in multiple countries. $PM_{2.5}$ concentrations showed spatial and temporal variations across the European sites in this study. Application of the statistical model with AOD in southeastern Europe indicated that this developed method could estimate $PM_{2.5}$ concentrations with reasonable accuracy at different spatial and temporal scales and provides a new approach for AOD-derived $PM_{2.5}$ estimation in the region. Finally, it should be noted that the model developed in this study reproduces the mean historical $PM_{2.5}$ concentrations, but since AOD is not an instantaneous online data set, the model cannot be used to predict future concentrations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11869-022-01245-5.

Acknowledgements The authors gratefully acknowledge NASA for making publicly available the VIIRS aerosol products and MERRA-2 data and the European Environment Agency (EEA) for $PM_{2.5}$ data.

Author contributions Serdar Gündogdu: formal analysis, methodology, investigation, and writing—original draft.

Gizem Tuna Tuygun: formal analysis; methodology; investigation; writing—original draft; writing—review and editing; and visualization.

Zhanqing Li: methodology; writing—original draft; and writing—review and editing.

Jing Wei: methodology; writing—original draft; and writing—review and editing.

Tolga Elbir: conceptualization; methodology; writing—review and editing; and supervision.

The authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability Data will be made available on reasonable request.

Declarations

Ethics approval This study did not require ethics approval.

Conflict of interest The authors declare no competing interests.

References

Achilleos S, Mouzourides P, Kalivitis N et al (2020). Spatio-temporal variability of desert dust storms in Eastern Mediterranean (Crete, Cyprus, Israel) between 2006 and 2017 using a uniform methodology. Sci Total Environ 714:136693. https://doi.org/10.1016/j.scitotenv.2020.136693

Adães J, Pires JCM (2019) Analysis and modelling of PM2.5 temporal and spatial behaviors in European cities. Sustain 11:6019. https://doi.org/10.3390/su11216019

Ahmad M, Alam K, Tariq S, Anwar S, Nasir J, Mansha M (2019) Estimating fine particulate concentration using a combined approach of linear regression and artificial neural network. Atmos Environ 219:117050. https://doi.org/10.1016/j.atmosenv.2019.117050

Almeida SM, Manousakas M, Diapouli E et al (2020) Ambient particulate matter source apportionment using receptor modelling in European and Central Asia urban areas. Environ Pollut 266: 115199. https://doi.org/10.1016/j.envpol.2020.115199

Barmpadimos I, Keller J, Oderbolz D et al (2012) One decade of parallel fine (PM 2.5) and coarse (PM 10-PM 2.5) particulate matter measurements in Europe: trends and variability. Atmos Chem Phys 12:3189–3203. https://doi.org/10.5194/acp-12-3189-2012

Barnes MJ, Brade TK, Mackenzie AR et al (2014) Spatially-varying surface roughness and ground-level air quality in an operational dispersion model. Environ Pollut 185:44–51. https://doi.org/10.1016/j.envpol.2013.09.039

Beloconi A, Chrysoulakis N, Lyapustin A et al (2018) Bayesian geostatistical modelling of PM10 and PM25 surface level concentrations in Europe using high-resolution satellite-derived products. Environ Int 121:57–70. https://doi.org/10.1016/j.envint.2018.08.041

Biecek P, Burzykowski T (2021) Explanatory model analysis: explore, explain, and examine predictive models. With examples in R and Python, New York

Carreño G, López-Cortés XA, Marchant C (2022) Machine learning models to predict critical episodes of environmental pollution for PM2.5 and PM10 in Talca, Chile. Mathematics 10(3). https://doi.org/10.3390/math10030373

Chen YW, Lin CJ (2006) Combining SVMs with various feature selection strategies. Stud Fuzziness Soft Comput 207:315–324. https://doi.org/10.1007/978-3-540-35488-8_13

Chen R, Gao Q, Sun J et al (2020) Short-term effects of particulate matter exposure on emergency room visits for cardiovascular disease in Lanzhou, China: a time series analysis. Environ Sci Pollut Res 27:9327–9335. https://doi.org/10.1007/s11356-020-07606-w

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016:785–794. https://doi.org/10.1145/2939672.2939785

Chen J, Yin J, Zang L et al (2019) Stacking machine learning model for estimating hourly PM2.5 in China based on Himawari 8 aerosol optical depth data. Sci Total Environ 697:134021. https://doi.org/10.1016/j.scitotenv.2019.134021

Choubin B, Abdolshahnejad M, Moradi E et al (2020) Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain. Sci Total Environ 701:134474. https://doi.org/10.1016/j.scitotenv.2019.134474

Czernecki B, Półrolniczak M, Kolendowicz L, Marosz M, Kendzierski S, Pilguj N (2017) Influence of the atmospheric conditions on PM10 concentrations in Poznań. Poland J Atmos Chem 74:115–139. https://doi.org/10.1007/s10874-016-9345-5

Czernecki B, Marosz M, Jędruszkiewicz J (2021) Assessment of machine learning algorithms in short-term forecasting of pm10 and pm2.5 concentrations in selected polish agglomerations. Aerosol Air Qual Res 21:200586. https://doi.org/10.4209/aaqr.200586

Dai H, Huang G, Zeng H, Zhou F (2022) PM2.5 volatility prediction by XGBoost-MLP based on GARCH models. J Clean Prod 356:131898. https://doi.org/10.1016/j.jclepro.2022.131898

Di Q, Kloog I, Koutrakis P et al (2016) Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. Environ Sci Technol 50:4712–4721. https://doi.org/10.1021/acs.est.5b06121

Diapouli E, Manousakas M, Vratolis S et al (2017) Evolution of air pollution source contributions over one decade, derived by PM10 and PM2.5 source apportionment in two metropolitan urban areas in Greece. Atmos Environ 164:416–430. https://doi.org/10.1016/j.atmosenv.2017.06.016

EEA (European Environment Agency) (2018) Air quality in Europe-2018 Report: EEA Report No 12/2018. https://www.eea.europa.eu/publications/air-quality-in-europe-2018. Accessed 20 May 2021

Feng L, Li Y, Wang Y, Du Q (2020) Estimating hourly and continuous ground-level PM2.5 concentrations using an ensemble learning algorithm: the ST-stacking model. Atmos Environ 223:117242. https://doi.org/10.1016/j.atmosenv.2019.117242

Gelaro R, McCarty W, Suárez MJ et al (2017) The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J Clim 30:5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Ghahremanloo M, Choi Y, Sayeed A et al (2021) Estimating daily high-resolution PM2.5 concentrations over Texas: machine learning approach. Atmos Environ 247:118209. https://doi.org/10.1016/j.atmosenv.2021.118209

Gui K, Che H, Zeng Z et al (2020) Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model. Environ Int 141:105801. https://doi.org/10.1016/j.envint.2020.105801

Guo J, Xia F, Zhang Y et al (2017) Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: ımplications for PM2.5 remote sensing. Environ Pollut 221:94–104. https://doi.org/10.1016/j.envpol.2016.11.043

Houthuijs D, Breugelmans O, Hoek G et al (2001) PM10 and PM2.5 concentrations in Central and Eastern Europe: results from the Cesar study. Atmos Environ 35:2757–2771. https://doi.org/10.1016/S1352-2310(01)00123-6

Hu X, Belle JH, Meng X et al (2017) Estimating PM2.5 Concentrations in the conterminous United States using the random forest approach. Environ Sci Technol 51:6936–6944. https://doi.org/10.1021/acs.est.7b01210

Huang K, Xiao Q, Meng X et al (2018) Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environ Pollut 242:675–683. https://doi.org/10.1016/j.envpol.2018.07.016

Ichoku C, Allen Chu D, Mattoo S et al (2002) A spatio-temporal approach for global validation and analysis of MODIS aerosol products. Geophys Res Lett 29:1616. https://doi.org/10.1029/2001GL013206

Jędruszkiewicz J, Czernecki B, Marosz M (2017) The variability of PM10 and PM2.5 concentrations in selected Polish agglomerations: the role of meteorological conditions, 2006–2016. Int J Environ Health Res 27(6):441–462. https://doi.org/10.1080/09603123.2017.1379055

Jiang T, Chen B, Nie Z et al (2021) Estimation of hourly full-coverage PM2.5 concentrations at 1-km resolution in China using a two-stage random forest model. Atmos Res 248:105146. https://doi.org/10.1016/j.atmosres.2020.105146

Jo EJ, Lee WS, Jo HY et al (2017) Effects of particulate matter on respiratory disease and the impact of meteorological factors in Busan, Korea. Respir Med 124:79–87. https://doi.org/10.1016/j.rmed.2017.02.010

Kaneyasu N, Yamamoto S, Sato K et al (2014) Impact of long-range transport of aerosols on the PM2.5 composition at a major metropolitan area in the northern Kyushu area of Japan. Atmos Environ 97:416–425. https://doi.org/10.1016/j.atmosenv.2014.01.029

Kaskaoutis DG, Rashki A, Dumka UC et al (2019) Atmospheric dynamics associated with exceptionally dusty conditions over the eastern Mediterranean and Greece in March 2018. Atmos Res 218:269–284. https://doi.org/10.1016/j.atmosres.2018.12.009

Khan R, Konishi S, Ng CFS et al (2019) Association between short-term exposure to fine particulate matter and daily emergency room visits at a cardiovascular hospital in Dhaka, Bangladesh. Sci Total Environ 646:1030–1036. https://doi.org/10.1016/j.scitotenv.2018.07.288

Levy RC, Munchak LA, Mattoo S et al (2015) Towards a long-term global aerosol optical depth record: applying a consistent aerosol retrieval algorithm to MODIS and VIIRS-observed reflectance. Atmos Meas Tech 8:4083–4110. https://doi.org/10.5194/amt-8-4083-2015

Li T, Shen H, Yuan Q, Zhang L (2018) Deep learning for ground-level PM2.5 prediction from satellite remote sensing data. In: International Geoscience and Remote Sensing Symposium (IGARSS), pp 7581–7584

Liu Y, Franklin M, Kahn R, Koutrakis P (2007) Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: a comparison between MISR and MODIS. Remote Sens Environ 107:33–44. https://doi.org/10.1016/j.rse.2006.05.022

Liu Y, Paciorek CJ, Koutrakis P (2009) Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information. Environ Health Perspect 117:886–892. https://doi.org/10.1289/ehp.0800123

Liu Q, Xu C, Ji G et al (2017) Effect of exposure to ambient PM2.5 pollution on the risk of respiratory tract diseases: a meta-analysis of cohort studies. J Biomed Res 31:130–142. https://doi.org/10.7555/JBR.31.20160071

Luo M, Wang Y, Xie Y, Zhou L, Qiao J, Qiu S, Sun Y (2021) Combination of feature selection and catboost for prediction: the first application to the estimation of aboveground biomass. Forests 12(2):1–22. https://doi.org/10.3390/f12020216

Lv L, Wei P, Li J, Hu J (2021) Application of machine learning algorithms to improve numerical simulation prediction of PM2.5 and chemical components. Atmos Pollut Res 12:101211. https://doi.org/10.1016/j.apr.2021.101211

Maciejewska K (2020) Short-term impact of PM2.5, PM10, and PMc on mortality and morbidity in the agglomeration of Warsaw. Poland Air Qual Atmos Heal 13:659–672. https://doi.org/10.1007/s11869-020-00831-9

Maji KJ, Dikshit AK, Arora M, Deshpande A (2018) Estimating premature mortality attributable to PM2.5 exposure and benefit of air pollution control policies in China for 2020. Sci Total Environ 612:683–693. https://doi.org/10.1016/j.scitotenv.2017.08.254

Meng X, Fu Q, Ma Z et al (2016) Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model. Environ Pollut 208:177–184. https://doi.org/10.1016/j.envpol.2015.09.042

Nabavi SO, Haimbergera L, Abbasib E (2019) Assessing PM2.5 concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms. Atmos Pollut Res 10:889–903. https://doi.org/10.1016/j.apr.2018.12.017

Năstase G, Șerban A, Năstase AF et al (2018) Air quality, primary air pollutants and ambient concentrations inventory for Romania. Atmos Environ 184:292–303. https://doi.org/10.1016/j.atmosenv.2018.04.034

NOAA (National Oceanic and Atmospheric Administration) (2020) SURFRAD aerosol optical depth. https://gml.noaa.gov/grad/surfrad/aod/. Accessed 2 Dec 2021

Ozdemir E, Tuna Tuygun G, Elbir T (2020) Application of aerosol classification methods based on AERONET version 3 product over eastern Mediterranean and Black Sea. Atmos Pollut Res 11:2226–2243. https://doi.org/10.1016/j.apr.2020.06.008

Park S, Shin M, Im J, Song CK, Choi M, Kim J, Lee S, Park R, Kim J, Lee DW, Kim SK (2019) Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea. Atmos Chem Phys 19(2):1097–1113. https://doi.org/10.5194/acp-19-1097-2019

Park S, Lee J, Im J et al (2020) Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models. Sci Total Environ 713:136516. https://doi.org/10.1016/j.scitotenv.2020.136516

Pražnikar ZJ, Pražnikar J (2012) The effects of particulate matter air pollution on respiratory health and on the cardiovascular system. Zdr Varst 51:190–199. https://sciendo.com/article/10.2478/v10152-012-0022-z

Randles CA, da Silva AM, Buchard V et al (2017) The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. J Clim 30:6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1

Sawyer V, Levy RC, Mattoo S et al (2020) Continuing the MODIS dark target aerosol time series with VIIRS. Remote Sens 12:308. https://doi.org/10.3390/rs12020308

Sayer AM, Hsu NC, Lee J et al (2018) Satellite ocean aerosol retrieval (SOAR) algorithm extension to S-NPP VIIRS as part of the "deep blue" aerosol project. J Geophys Res Atmos 123:380–400. https://doi.org/10.1002/2017JD027412

Shtein A, Kloog I, Schwartz J et al (2019) Estimating daily PM2.5 and PM10 over Italy using an ensemble model. Environ Sci Technol. https://doi.org/10.1021/acs.est.9b04279

Song P, Liu Y (2020) An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms. Teh Vjesn 27:1467–1471. https://doi.org/10.17559/TV-20200808113807

Stafoggia M, Schwartz J, Badaloni C et al (2017) Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ Int 99:234–244. https://doi.org/10.1016/j.envint.2016.11.024

Stafoggia M, Bellander T, Bucci S et al (2019) Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ Int 124:170–179. https://doi.org/10.1016/j.envint.2019.01.016

Tuna Tuygun G, Elbir T (2020) Long-term temporal analysis of the columnar and surface aerosol relationship with planetary boundary layer height at a southern coastal site of Turkey. Atmos Pollut Res 11:2259–2269. https://doi.org/10.1016/j.apr.2020.09.008

Tuna Tuygun G, Gündoğdu S, Elbir T (2021) Estimation of ground-level particulate matter concentrations based on synergistic use of MODIS, MERRA-2 and AERONET AODs over a coastal site in the Eastern Mediterranean. Atmos Environ 261:118562. https://doi.org/10.1016/j.atmosenv.2021.118562

Wang L, Liu Z, Sun Y et al (2015) Long-range transport and regional sources of PM2.5 in Beijing based on long-term observations from 2005 to 2010. Atmos Res 157:37–48. https://doi.org/10.1016/j.atmosres.2014.12.003

Wang Y, Yuan Q, Li T et al (2021) Full-coverage spatiotemporal mapping of ambient PM2.5 and PM10 over China from Sentinel-5P and assimilated datasets: considering the precursors and chemical compositions. Sci Total Environ 793:148535. https://doi.org/10.1016/j.scitotenv.2021.148535

Wei J, Li Z, Cribb M et al (2020) Improved 1&thinsp;km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees. Atmos Chem Phys 20:3273–3289. https://doi.org/10.5194/acp-20-3273-2020

Wei J, Li Z, Sun L, et al (2021) Extending the EOS long-term PM2.5Data Records since 2013 in China: application to the VIIRS Deep Blue Aerosol Products. IEEE Trans Geosci Remote Sens 60. https://doi.org/10.1109/TGRS.2021.3050999

Wei J, Huang W, Li Z et al (2019) Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens Environ 231:111221. https://doi.org/10.1016/j.rse.2019.111221

Wu J, Yao F, Li W, Si M (2016) VIIRS-based remote sensing estimation of ground-level PM2.5 concentrations in Beijing–Tianjin–Hebei: a spatiotemporal statistical model. Remote Sens Environ 184:316–328. https://doi.org/10.1016/j.rse.2016.07.015

Xiao Q, Wang Y, Chang HH et al (2017) Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sens Environ 199:437–446. https://doi.org/10.1016/j.rse.2017.07.023

Xing YF, Xu YH, Shi MH, Lian YX (2016) The impact of PM2.5 on the human respiratory system. J Thorac Dis 8:E69–E74. https://jtd.amegroups.com/article/view/6353/6196

Xue T, Zheng Y, Tong D et al (2019) Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ Int 123:345–357. https://doi.org/10.1016/j.envint.2018.11.075

Yan X, Zang Z, Luo N et al (2020) New interpretable deep learning model to monitor real-time PM2.5 concentrations from satellite data. Environ Int 144:106060. https://doi.org/10.1016/j.envint.2020.106060

Yan X, Zang Z, Jiang Y et al (2021) A spatial-temporal ınterpretable deep learning model for improving interpretability and predictive accuracy of satellite-based PM2.5. Environ Pollut 273:116459. https://doi.org/10.1016/j.envpol.2021.116459

Yao F, Si M, Li W, Wu J (2018) A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. Sci Total Environ 618:819–828. https://doi.org/10.1016/j.scitotenv.2017.08.209

Yao F, Wu J, Li W, Peng J (2019) A spatially structured adaptive two-stage model for retrieving ground-level PM 2.5 concentrations from VIIRS AOD in China. ISPRS J Photogramm Remote Sens 151:263–276. https://doi.org/10.1016/j.isprsjprs.2019.03.011

Yazdi MD, Kuang Z, Dimakopoulou K et al (2020) Predicting fine particulate matter (PM2.5) in the greater london area: an ensemble approach using machine learning methods. Remote Sens 12:914. https://doi.org/10.3390/rs12060914

Yue W, Tong L, Liu X et al (2019) Short term PM2.5 exposure caused a robust lung inflammation, vascular remodeling, and exacerbated transition from left ventricular failure to right ventricular hypertrophy. Redox Biol 22:101161. https://doi.org/10.1016/j.redox.2019.101161

Zeydan Ö, Wang Y (2019) Using MODIS derived aerosol optical depth to estimate ground-level PM2.5 concentrations over Turkey. Atmos Pollut Res 10:1565–1576. https://doi.org/10.1016/j.apr.2019.05.005

Zhang X, Chu Y, Wang Y, Zhang K (2018) Predicting daily PM2.5 concentrations in Texas using high-resolution satellite aerosol optical depth. Sci Total Environ 631–632:904–911. https://doi.org/10.1016/j.scitotenv.2018.02.255

Zhang G, Lu H, Dong J, Poslad S, Li R, Zhang X, Rui X (2020) A framework to predict high-resolution spatiotemporal pm25 distributions using a deep-learning model: a case study of shijiazhuang, china. Remote Sens 12(17):1–33. https://doi.org/10.3390/rs12172825

Zhang T, He W, Zheng H et al (2021) Satellite-based ground PM2.5 estimation using a gradient boosting decision tree. Chemosphere 268:128801. https://doi.org/10.1016/j.chemosphere.2020.128801

Zhao C, Liu Z, Wang Q, Ban J, Chen NX, Li T (2019) High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing-Tianjin-Hebei region. Atmos Environ 203:70–78. https://doi.org/10.1016/j.atmosenv.2019.01.045

Zhao C, Wang Q, Ban J et al (2020) Estimating the daily PM2.5 concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01° × 0.01° spatial resolution. Environ Int 134:105297. https://doi.org/10.1016/j.envint.2019.105297

Zheng H, Wu Y (2019) A XGBoost model with weather similarity analysis and feature engineering for short-term wind power forecasting. Appl Sci 9:3019. https://doi.org/10.3390/app9153019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.