remote sensing

MDPI

Technical Note

How Important Is Satellite-Retrieved Aerosol Optical Depth in Deriving Surface $PM_{2.5}$ Using Machine Learning?

Zhongyan Tian [1,†], Jing Wei [2,†] and Zhanqing Li [2,*]

[1] Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; 202021490035@mail.bnu.edu.cn
[2] Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20740, USA; weijing@umd.edu
* Correspondence: zhanqing@umd.edu
† These authors contributed equally to this work.

Abstract: $PM_{2.5}$ refers to the total mass concentration of tiny particulates in the atmosphere near the surface, obtained by means of in situ observations and satellite remote sensing. Given the highly limited number of ground observation stations of inhomogeneous distribution and an ill-posed remote sensing approach, increasing efforts have been devoted to the application of machine-learning (ML) models to both ground and satellite data. A key satellite-derived parameter, aerosol optical thickness (AOD), has been most commonly used as a proxy of $PM_{2.5}$, although their correlation is fraught with large uncertainties. A critical question that has been overlooked concerns how much AOD helps to improve the retrieval of $PM_{2.5}$ relative to its uncertainty incurred concurrently. The question is addressed here by taking advantage of high-density $PM_{2.5}$ stations in eastern China to evaluate the contributions of AOD, determined as the difference in the accuracy of $PM_{2.5}$ retrievals with and without AOD for varying densities of $PM_{2.5}$ stations, using four popular ML models (i.e., Random Forest, Extra-trees, XGBoost, and LightGBM). Our results reveal that as the density of monitoring stations decreases, both the feature importance and permutation importance of satellite AOD demonstrate a consistent upward trend ($p < 0.05$). Furthermore, the ML models without AOD exhibit faster declines in overall accuracy and predictive ability compared with the models with AOD assessed using the sample-based and station-based (spatial) independent cross-validation approaches. Overall, a 10% reduction in the number of stations results in an increase of 0.7–1.2% and 0.6–1.2% in uncertainty in estimated and predicted accuracies, respectively. These findings attest to the indispensable role of satellite AOD in the $PM_{2.5}$ retrieval process through ML because it can significantly mitigate the negative impact of the sparse distribution of monitoring sites. This role becomes more important as the number of $PM_{2.5}$ stations decreases.

Keywords: machine learning; AOD; $PM_{2.5}$ retrieval; station density; importance assessment

Citation: Tian, Z.; Wei, J.; Li, Z. How Important Is Satellite-Retrieved Aerosol Optical Depth in Deriving Surface $PM_{2.5}$ Using Machine Learning? *Remote Sens.* **2023**, *15*, 3780. https://doi.org/10.3390/rs15153780

Academic Editor: Stephan Havemann

Received: 30 June 2023
Revised: 26 July 2023
Accepted: 26 July 2023
Published: 29 July 2023

1. Introduction

$PM_{2.5}$ refers to the concentration of airborne particulate matter (PM) with an aerodynamic diameter of less than 2.5 microns. Although small, these particles are abundant and active, and attach easily to toxic and harmful substances. $PM_{2.5}$ can be suspended in the atmosphere for extended periods, ranging from months to even years, which has an important impact on air quality and visibility and also affects human health [1–3]. While $PM_{2.5}$ has been monitored in many parts of the world, observations are still highly limited and very inhomogeneous, with many regions not covered [4–6]. However, satellite remote sensing provides continuous spatial coverage and has been widely used in the estimation of surface $PM_{2.5}$ concentrations [7–9].

Previous studies have made great efforts to infer $PM_{2.5}$ from satellite retrievals of aerosol optical depth (AOD) by virtue of their positive correlation because AOD is much

more monitored from both space and the ground. Many factors can influence their relationship, including aerosol vertical distribution, relative humidity, mixed-layer height, and topography, among others [10–12]. The relationship also varies with both location and time scale. Wang and Christopher (2003) [13] used the AOD product retrieved by the Moderate Resolution Imaging Spectroradiometer (MODIS) and in situ measurements of $PM_{2.5}$ at seven ground observation stations in Alabama, USA, finding a sound correlation between them on a monthly time scale. Natunen et al. (2010) [14] explored the relationship at four stations in Helsinki, Finland, on seasonal and monthly time scales and found that time averaging increased the correlation. Likewise, the correlation also varies with spatial resolution [15], indicating that different geographical locations, study area sizes, and spatial resolutions of MODIS AOD products can change the correlation between AOD and $PM_{2.5}$. In general, the relationship varies considerably with location and season [16,17]. Su et al. (2018) [18] studied the relationship extensively across China, one of the most polluted regions of the world, finding that the relationship differs considerably in different parts of China (better in northern than in southern China) and among the four seasons (better in winter than in summer). The relationship can be significantly improved by normalizing against the height of the planetary boundary layer.

Due to the complex relationships between AOD and $PM_{2.5}$, many statistical regression methods have been proposed for estimating $PM_{2.5}$ using satellite AOD retrievals [7,19–21], such as the multiple linear regression model [22], the geographically weighted regression model [23,24], the geographical spatiotemporal weighting regression model [25], and the linear mixed effects model [26]. To a certain extent, these models are capable of estimating surface $PM_{2.5}$ concentrations using satellite AOD data. However, they face challenges when it comes to studying the influences of various factors on $PM_{2.5}$, such as meteorological factors (boundary layer height, relative humidity, etc.) and surface factors (underlying surface types, etc.) [27,28]. Fortunately, machine-learning (ML) models have a strong data-mining capability and can establish robust nonlinear relationships. They allow for the extraction of pertinent information from very large numbers of auxiliary factors to improve the accuracy of $PM_{2.5}$ retrievals. Therefore, various types of ML models have been adopted in $PM_{2.5}$ inversion studies in recent years, e.g., the Random Forest model [29,30], the Extra-trees model [31,32], the XGBoost model [33], and the LightGBM model [34].

AOD has been regarded as an essential input variable in inferring $PM_{2.5}$ from satellites [7,19,27,35]. However, a handful of studies have presented contrasting results [36–39]. Chen et al. (2021) [38], for example, developed a Random Forest model for areas with and without AOD data, finding that the model or areas without AOD can result in better $PM_{2.5}$ retrievals. Yu et al. (2022) [39] developed a deep ensemble ML framework to estimate daily $PM_{2.5}$ concentrations in Italy from 2015 to 2019 and found similar accuracies (cross-validated R^2 = 0.853 and 0.857) in comparison with ground observations when including or not including satellite AOD in the model. These conflicting findings pose such critical questions as whether satellite-retrieved AOD plays any significant role in estimating surface $PM_{2.5}$ and what factor, if any, dictates its role in the ML application for estimating $PM_{2.5}$. We attempt to address these questions by taking advantage of rich satellite AOD data and in situ $PM_{2.5}$ measurements in eastern China, together with a large array of other ancillary data, introduced next.

2. Data and Methods

2.1. Study Area

The study area (approximately 1,830,000 km^2) covers 14 provinces in China, including the North China Plain, the Yangtze River Delta, the Pearl River Delta, and parts of central China (Figure 1). As the most populated and advanced in economic development in China, these regions have experienced serious air pollution problems, thus garnering significant public attention. To monitor air pollution, relatively dense $PM_{2.5}$ ground observation stations have been uniformly distributed, enabling us to investigate the effects of

satellite AOD on estimating $PM_{2.5}$ concentrations at different levels of ground observation station density.

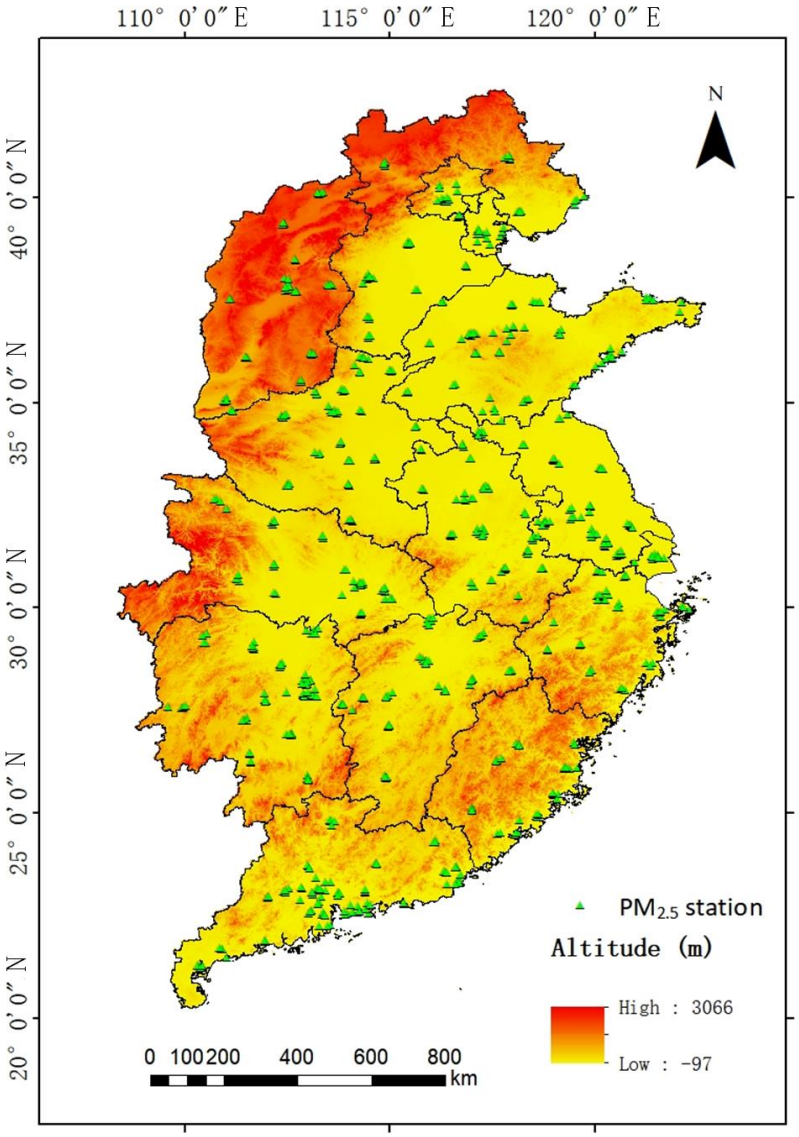

Figure 1. Study area and the distribution of ground stations (green triangles). The colored background shows land elevations (unit: m).

2.2. Data Sources

The datasets used in this study consist of observed $PM_{2.5}$ concentrations, 1-km-resolution MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD products, and many auxiliary datasets related to $PM_{2.5}$, such as meteorological and land- and population-related information. The study period spans from 2018 to 2020, ensuring an adequate volume of data for conducting sensitivity analyses.

2.2.1. $PM_{2.5}$ Ground Measurements

The $PM_{2.5}$ observation station data used in this study are real-time ground-measured air quality data (including $PM_{2.5}$) in China from the China National Environment Monitoring Center. In this study, a total of 775 ground observation stations were chosen from the eastern region of China (Figure 1). Daily measurements recorded at each station were subsequently calculated and subjected to rigorous data quality control following Wei et al.

(2019) [30]. These validated ground measurements were then utilized as ground truth for ML-modeling purposes.

2.2.2. MODIS AOD Products

The MODIS AOD product serves as the primary predictor for estimating surface $PM_{2.5}$ in this study. Specifically, Terra and Aqua MCD19A2 AOD products at a spatial resolution of 1 km are employed. This product is retrieved using the MAIAC inversion algorithm over land and incorporates various quality assurance (QA) measures [40,41]. For this study, to ensure the quality of the data, we only employed those MAIAC AOD retrievals passing the recommended QA measures, including cloud screening (QACloudMask = Clear) and adjacency (QAAdjacencyMask = Clear), following the methodology outlined in our previous study [30].

2.2.3. Auxiliary Data

Meteorological reanalysis data used in this paper are collected from the fifth-generation European Reanalysis Interim dataset (ERA5) released by the European Centre for Medium-Range Weather Forecasts. The global hourly dataset has characterized the states of the atmosphere, oceans, and surface since 1979 [42]. Specifically, seven meteorological parameters were employed: boundary layer height (BLH; unit: m), evaporation (ET; unit: mm), relative humidity (RH; unit: %), surface pressure (SP; unit: hPa), 2 m air temperature (TEM; unit: K), and 10 m U and V wind components (unit: m s^{-1}). Copernicus Atmospheric Monitoring Service (CAMS) emission inventories, including the four main precursors of $PM_{2.5}$, i.e., ammonia, nitrogen oxides, sulfur dioxide, and volatile organic compounds, were also considered [43,44]. In addition, parameters related to surface conditions and human activities, including the normalized vegetation index, a digital elevation model, and population density, were involved. In total, 15 predictor variables, including AOD, are utilized for $PM_{2.5}$ modeling through ML.

2.3. Methodology

2.3.1. Machine-Learning (ML) Models

ML applies complex statistical theories and algorithms to computer simulations in a somewhat human-learning behavior to acquire new knowledge. It can take advantage of any existing knowledge structure and ample information content to infer a piece of new knowledge. ML has been applied in various fields, including the remote sensing of $PM_{2.5}$ [45–47]. This study compares and analyzes four popular widely used ML algorithms, i.e., Random Forest, Extra-trees, XGBoost, and LightGBM models.

The Random Forest model constructs an ensemble of multiple decision trees, where each tree is generated by bootstrap sampling from the training dataset [48–50]. The Extremely Randomized Trees (Extra-trees) model is also a tree-based ensemble learning method, similar to Random Forest but introduces additional randomness in selecting features and splitting the points from all data samples in the tree-building process [31,32,51]. Both the Random Forest and Extra-trees models are ensemble-learning algorithms based on the bagging technique. In the bagging training process, the base classifiers (decision trees) are trained independently, and there is no strong dependence or correlation between them. This characteristic allows for parallel training of the base classifiers, which can significantly speed up the training process. By training the base classifiers in parallel, these integrated algorithms harness the power of parallel computing, making them efficient and scalable for large datasets.

In contrast, the XGBoost and LightGBM models are based on a boosting ensemble algorithm (Figure 2), where base classifiers are trained sequentially, and each classifier depends on the others. The main goal of boosting is to stack these classifiers on top of each other, with each layer assigning higher weights to samples that were incorrectly classified by the previous layers. However, these two models differ in several ways. XGBoost utilizes a pre-classification algorithm, meaning that all features of a sample are pre-sorted before

iterated and repeated operations take place [33,52]. This sorting step significantly reduces the number of calculations required. LightGBM employs a histogram algorithm, which offers advantages such as reduced memory usage and a faster runtime [34,53]. In terms of growth strategy, XGBoost follows a level-wise approach. In this strategy, the child nodes of the same layer are split simultaneously. Conversely, LightGBM adopts a leaf-wise growth strategy, where each layer's child node only needs to find the node with the largest gain (typically the one with the largest data volume) to perform the split.

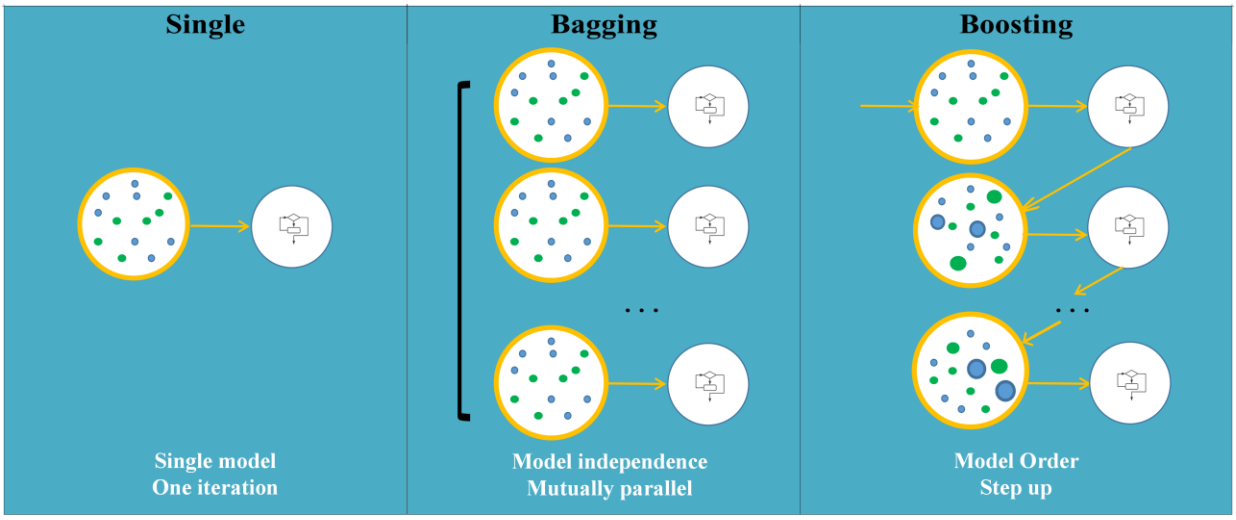

Figure 2. Illustration of the bagging and boosting algorithms.

2.3.2. Importance Assessment Method

(1) Feature importance

The feature importance (FI) score is a common indicator reflecting the importance of input variables that comes with the tree-based ML models. The FI score is calculated via the Gini index based on the mean decrease impurity (MDI) and used to evaluate the importance of each feature by measuring its contribution to splitting in the decision tree [54]. Taking one decision tree as an example, VIM represents variable importance measures, and GI represents the Gini index. Assuming that there are m feature variables, the GI score of a feature (represented by X_i) is calculated (represented by VIM_i^{Gini}), i.e., the average change in the node-splitting impurity of the ith feature in the tree model. The Gini index is calculated as

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad (1)$$

where K represents the total number of categories of a feature, and p_{mk} represents the proportion of a category k in node m. The importance of feature X_i in node m, i.e., the GI change before and after the node branch is expressed as

$$VIM_{im}^{(Gini)} = GI_m - GI_l - GI_r \quad (2)$$

where GI_l and GI_r represent the Gini indices of the two new nodes after the branch. The node where feature X_i appears in the jth decision tree is set M. The importance of X_i in the jth tree then is

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{im}^{(Gini)} \quad (3)$$

(2) Permutation importance

Although the FI can reflect the characteristic importance of variables, it is more favorable when there are more variable categories. For characteristic variables with multiple

correlations, FI may have a bias in describing correlation features, and its assessment could also be overfitted [55]. Therefore, the permutation importance (PI), another method for evaluating the contribution of each feature, is employed. The PI is a model-independent method applicable to almost all types of models, including deep-learning models. Its basic idea is to evaluate the importance of features on a test set by randomly shuffling a feature and then measuring the change in model performance to measure the importance of features. The PI of a feature is calculated as follows (Figure 3). First, the dataset is divided into a training set and a validation set. Second, a baseline metric, defined by scoring, is then evaluated on a (potentially different) dataset defined by the training set. Third, a feature column from the validation set is permutated randomly, and the metric is evaluated again. The PI can be obtained by calculating the difference between the baseline metric and the metric from permutating the feature column.

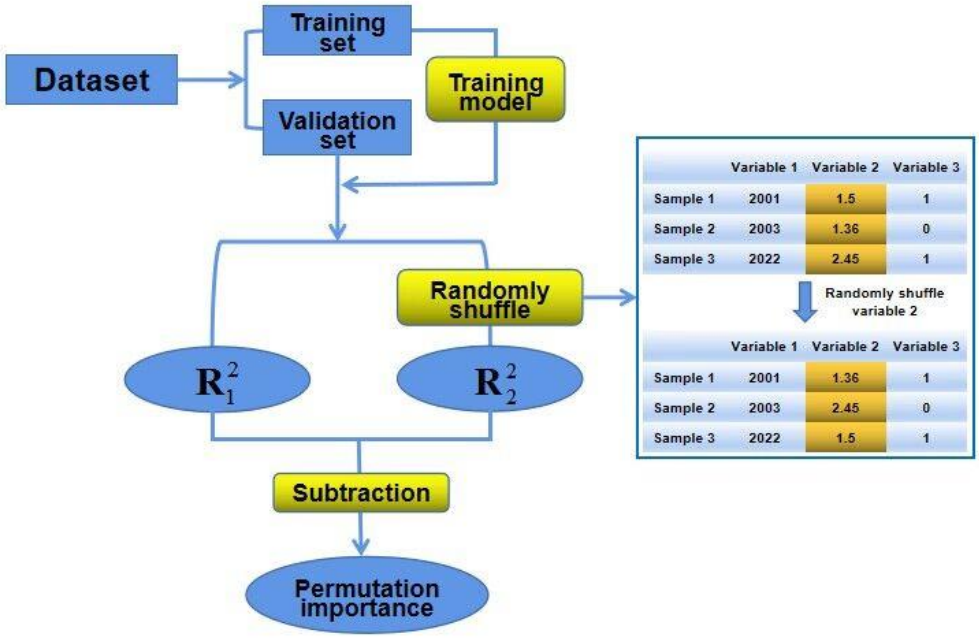

Figure 3. Illustration of the permutation importance (PI).

2.3.3. Model Validation Methods

In this study, two methods are used to evaluate the performance of ML models: sample-based and station-based ten-fold cross-validation (10-CV) [32,56]. In sample-based 10-CV, the sample dataset is randomly divided into ten groups. One group, comprising 10% of the samples, is set aside as the independent validation set, while the remaining nine groups (90%) form the training set. This process is repeated ten times, with each group serving as the validation set once, ensuring that all samples have been tested. The final accuracy is calculated as the average of the results obtained from the ten runs. This approach is commonly used to represent the overall accuracy of ML models in estimating $PM_{2.5}$ levels at locations where ground measurements are available.

The station-based 10-CV is another evaluation method used to assess the predictive ability of ML models in estimating $PM_{2.5}$ concentrations at locations where ground measurements are not available [57]. This method serves as a spatially independent validation technique. Similarly to the sample-based 10-CV, the station-based 10-CV involves dividing the ground observation stations in the study area into ten groups. One group, consisting of 10% of all stations, is designated as the validation set, while the dataset corresponding to the remaining nine groups of stations (90% of all stations) is used as the training set. This approach creates training and validation samples from different locations. This helps isolate spatial autocorrelations among the data samples, making it an effective spatially independent verification method.

2.3.4. Sensitivity Analysis Methods

This study starts by considering the study area as a whole, with 775 ground observation stations. The total number of stations in the study area is then randomly reduced by 10%. The remaining number of stations is again reduced by 10%. This process continues until the number of stations in a group is 31% of the original total number of stations. The end result is 12 groups of stations, each reflecting a certain station density in the study area. The smaller the proportion of remaining stations in a group, the smaller the density of stations in that group. This set of 12 groups of stations and their associated observations is used next to explore the influence of AOD on inversions of $PM_{2.5}$ from different ML models. To assess the importance and contribution of satellite AOD retrievals to ML modeling and quantitatively evaluate its impact on the performance of an ML model, this study incorporates an uncertainty analysis from three key aspects:

(1) The importance scores of satellite AOD were first calculated employing two techniques (FI and PI) for four typical tree-based ML models as the density of ground-based stations in the study area gradually decreased. This analysis offers valuable insights into the role of satellite AOD in the modeling process. It allows for an understanding of the significance of satellite AOD under different station-density conditions.
(2) The accuracies and differences in the estimation of $PM_{2.5}$, with and without satellite AOD as the primary predictor, were calculated using the sample-based 10-CV method. Four typical tree-based ML models were employed, each taking into consideration the decreasing density of ground-based stations in the study area. This analysis allows us to evaluate the importance of satellite AOD in enhancing the overall accuracy of $PM_{2.5}$ estimates for varying station densities.
(3) Similarly, the accuracies and differences in the prediction of $PM_{2.5}$ in regions lacking $PM_{2.5}$ observations, with and without satellite AOD as the main predictor, were calculated using the station-based 10-CV method. Again, four typical tree-based ML models were employed, each taking into consideration the decreasing density of ground-based stations in the study area. This analysis enables us to assess the significance of satellite AOD in improving the predictive ability of $PM_{2.5}$ predictions for varying station densities.

3. Results

3.1. Variations of Satellite AOD Contributions

Figure 4 presents the FI and PI of satellite AOD retrievals in estimating $PM_{2.5}$ obtained by four ML models with the decrease in the density of ground-based monitoring stations. Also shown are the best-fit lines from linear regression and confidence intervals (CIs). CI is a statistical measure used to quantify the uncertainty of an estimate. Here, we used a 95% confidence level (pink-shaded areas in the figure). This indicates that there is a 95% probability that the true value will fall within the specified range, leaving a 5% probability of it falling outside this range. Figure 4a–c show that as the density of the monitoring station decreases, the FI score significantly increases, with regressed correlation of determination (R^2) values of 0.84, 0.81, 0.77, and 0.43 for the Random Forest, Extra-trees, XGBoost, and LightGBM models, respectively. All pass the 99% or 95% confidence ($p < 0.01$ or 0.05) test. The spread of FI in the LightGBM case (Figure 4d) indicates a higher variance compared with other methods, which may be attributed to the use of the specific node-splitting method of the leaf-wise growth strategy. Similar conclusions can be made from the PI score analysis, i.e., the contribution of satellite AOD significantly increases as the density of monitoring stations decreases, with much higher regressed R^2 values (R^2 = 0.94–0.96) for the four ML models (Figure 4e–h). All regressed trends reach the 99% confidence ($p < 0.01$) level. Note that the values of FI and PI are different among the ML methods because these models employ distinct frameworks and operation methods, including sampling, feature selection, and node-splitting techniques. Additionally, the methods of computing FI and PI are also different, e.g., FI relies on the Gini index, which involves calculating the MDI, while PI assesses the change in model performance by randomly shuffling a feature. Results

obtained by the two importance verification methods are consistent, suggesting that the two methods can complement and verify each other. These findings reveal that satellite AOD is crucial for $PM_{2.5}$ modeling using ML because it plays a dominant predictive role with the highest importance scores, particularly in regions with a small density of $PM_{2.5}$ ground observation stations.

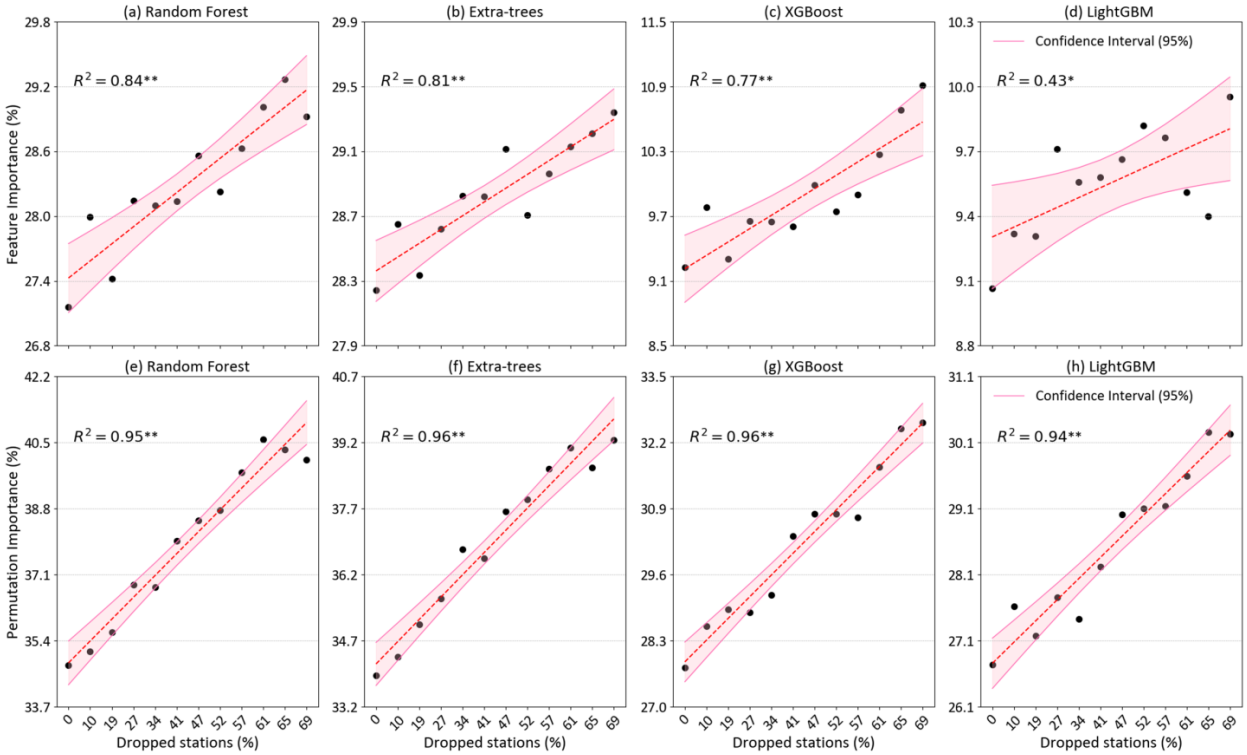

Figure 4. Variations of (**a**–**d**) feature importance (unit: %) and (**e**–**h**) permutation importance (unit: %) of satellite AOD as a function of the decreasing percentage of ground-based monitoring stations for the Random Forest, Extra-trees, XGBoost, and LightGBM models, respectively. Pink-shaded areas are 95% confidence intervals, where * and ** denote the 95% ($p < 0.05$) and 99% ($p < 0.01$) confidence levels of the regressed fits, respectively.

3.2. Impacts of Satellite AOD on Overall Accuracy

Figure 5 shows the overall accuracies measured by the CV R^2 (CV-R^2) of the daily estimates of $PM_{2.5}$ with and without AOD as model input as a function of decreasing station density using four ML models. Regarding model results where AOD was included as an input variable (Figure 5, red dots), it is clearly seen that as the station density decreases, the overall accuracy of the $PM_{2.5}$ estimates gradually decreases, showing an average significantly decreasing trend (i.e., change in sample-based CV-R^2 per 1% of discarded stations) of −0.16% ($p < 0.01$), −0.15% ($p < 0.01$), −0.19% ($p < 0.01$), and −0.11% ($p < 0.01$) for the Random Forest, Extra-trees, XGBoost, and LightGBM models, respectively. For model results where AOD was not included as an input variable (Figure 5, blue dots), when the station proportion is 100%, the CV-R^2 is 0.83 with AOD and 0.82 without AOD, a small difference of about 1.48%. However, as the station density decreases, the difference in accuracy between the models with and without AOD gradually increases (black diamonds in the figure). When the station density is 31%, the accuracy of the model with AOD is 0.72, while the accuracy of the model without AOD is 0.65. The relative difference in model accuracy with or without AOD significantly increases to 10.35%. Similarly, Figure 5b–d present comparable results: At 100% station density, the overall accuracies of the Extra-trees, XGBoost, and LightGBM models with AOD are 0.86, 0.84, and 0.85, respectively. Without AOD, the accuracies of these models are 0.86, 0.83, and 0.81, respectively, and

the relative differences in model accuracy with or without AOD are 0.10%, 0.58%, and 4.80%, respectively. When the station density drops to 31%, the accuracies of the models with AOD become 0.76, 0.70, and 0.76, respectively, while the accuracies of the models without AOD are 0.71, 0.64, and 0.69, respectively. Consequently, the relative differences in overall accuracy between these three models with and without AOD increase to 6.48%, 9.41%, and 10.44%, respectively. In particular, when the number of monitoring stations decreases, the slopes of the decreased overall accuracy are much steeper for the Random Forest (−0.24%, $p < 0.01$), Extra-trees (−0.22%, $p < 0.01$), XGBoost (−0.27%, $p < 0.01$), and LightGBM (−0.16%, $p < 0.01$) models without using AOD compared to these models using AOD. This is because with the decrease in station density, the sample data volume gradually decreases, as do the accuracies of the models. In general, for every 10% reduction in station proportion, the four ML models without AOD experience a 1.2% ($p < 0.01$), 0.9% ($p < 0.01$), 1.1% ($p < 0.01$), and 0.7% ($p < 0.01$) increase in the uncertainty of the estimated results, respectively, compared with these models using AOD. These findings highlight the indispensable role of satellite AOD in improving the accuracy of estimating $PM_{2.5}$ through ML models, particularly in regions with limited observation stations.

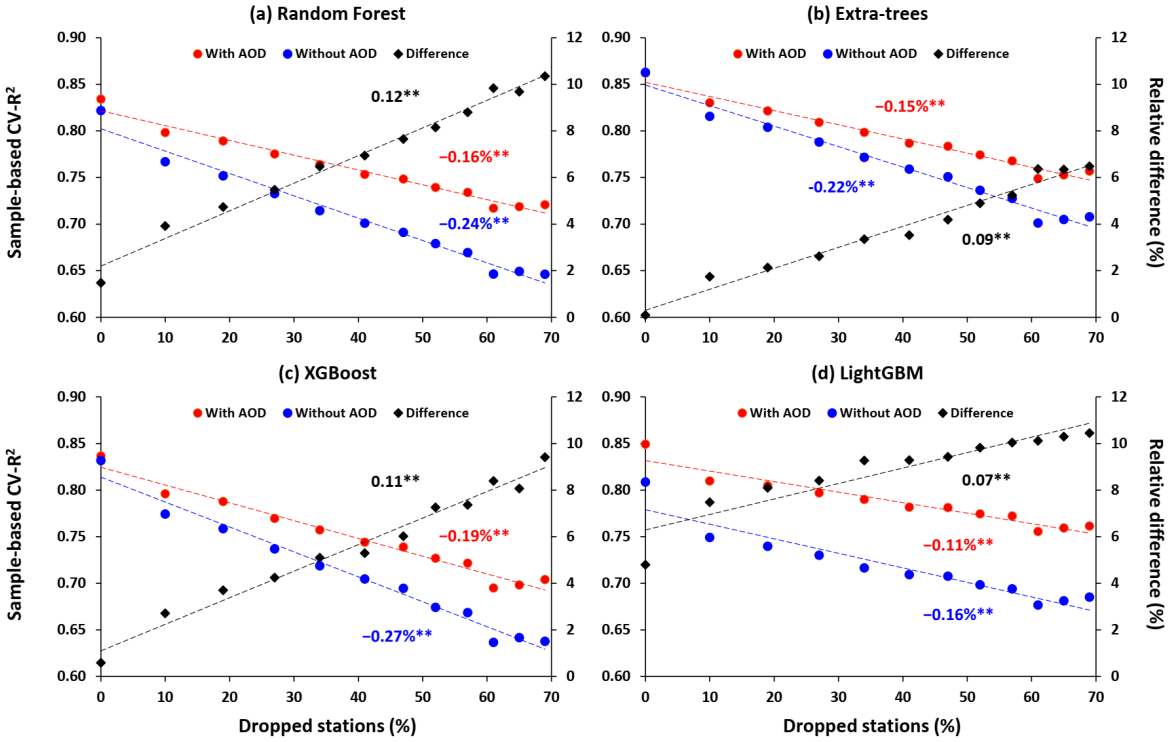

Figure 5. Variation in the overall accuracy (sample-based CV-R^2, left ordinate) and relative difference (%, right ordinate) of daily $PM_{2.5}$ estimates as a function of decreasing station density for four ML models: (**a**) Random Forest, (**b**) Extra-trees, (**c**) XGBoost, and (**d**) LightGBM, with (red dots) and without (blue dots) including satellite AOD as an input predictor. Slopes of the best-fit lines from linear regression for each set of results are given, where ** denotes the 99% confidence ($p < 0.01$) level. Black dashed lines represent the regressed fits of the relative difference between each set of results.

3.3. Impacts of Satellite AOD on Predictive Ability

Here, the model performance in predicting the $PM_{2.5}$ level in areas without surface observations is examined based on results from the station-based 10-CV method. Figure 6 shows the predictive ability of four ML models in retrieving daily $PM_{2.5}$ with (red dots) and without (blue dots) AOD as inputs under different station density conditions. Overall, the trend observed in station-based 10-CV results aligns with those of sample-based 10-CV: When AOD is included as an input variable, with the decreasing number of monitoring stations, the predictive accuracies of the Random Forest, Extra-trees, XGBoost, and Light-

GBM models all show significantly decreasing trends (i.e., change in station-based CV-R^2 per 1% of discarded stations) of −0.14% ($p < 0.01$), −0.13% ($p < 0.01$), −0.18% ($p < 0.01$), and −0.10% ($p < 0.01$), respectively. However, the model's predictive ability experiences a much faster decline, with larger slopes of −0.21%, −0.18%, −0.26%, and −0.13% for the four models without including satellite AOD, respectively. Specifically, when data from all $PM_{2.5}$ stations are used (100%), the Random Forest, Extra-trees, XGBoost, and LightGBM models that include AOD produce slightly better results, with higher station-based CV-R^2 values (0.81, 0.83, 0.80, and 0.79) than those without considering AOD (CV-R^2 = 0.77, 0.81, 0.76, and 0.71, respectively). However, when the proportion of monitoring stations drops to 31%, the predictive accuracies for the same four models decrease to 0.72, 0.75, 0.69, 0.73 (including AOD) and 0.64, 0.69, 0.61, 0.64 (not including AOD), respectively. More importantly, the average relative differences become 2.48, 2.68, 2.19, and 1.33 times larger than the results when all stations are considered. This is because station-based 10-CV uses known stations (regions) to predict unknown stations (regions), reflecting the spatial prediction ability of the model. However, it is difficult for an ML model to make predictions for regions without training samples, inevitably leading to lower prediction accuracies, consistent with existing cognitive rules of an ML model. In general, for every 10% reduction in the station density, the four ML models without and with AOD experience 1.0% ($p < 0.01$), 0.7% ($p < 0.01$), 1.2% ($p < 0.01$), and 0.6% ($p < 0.01$) increases in the uncertainty of the predicted results, respectively. These findings confirm the indispensable role of satellite AOD in predicting $PM_{2.5}$ concentration in areas without observations using ML models, particularly in low-station density situations.

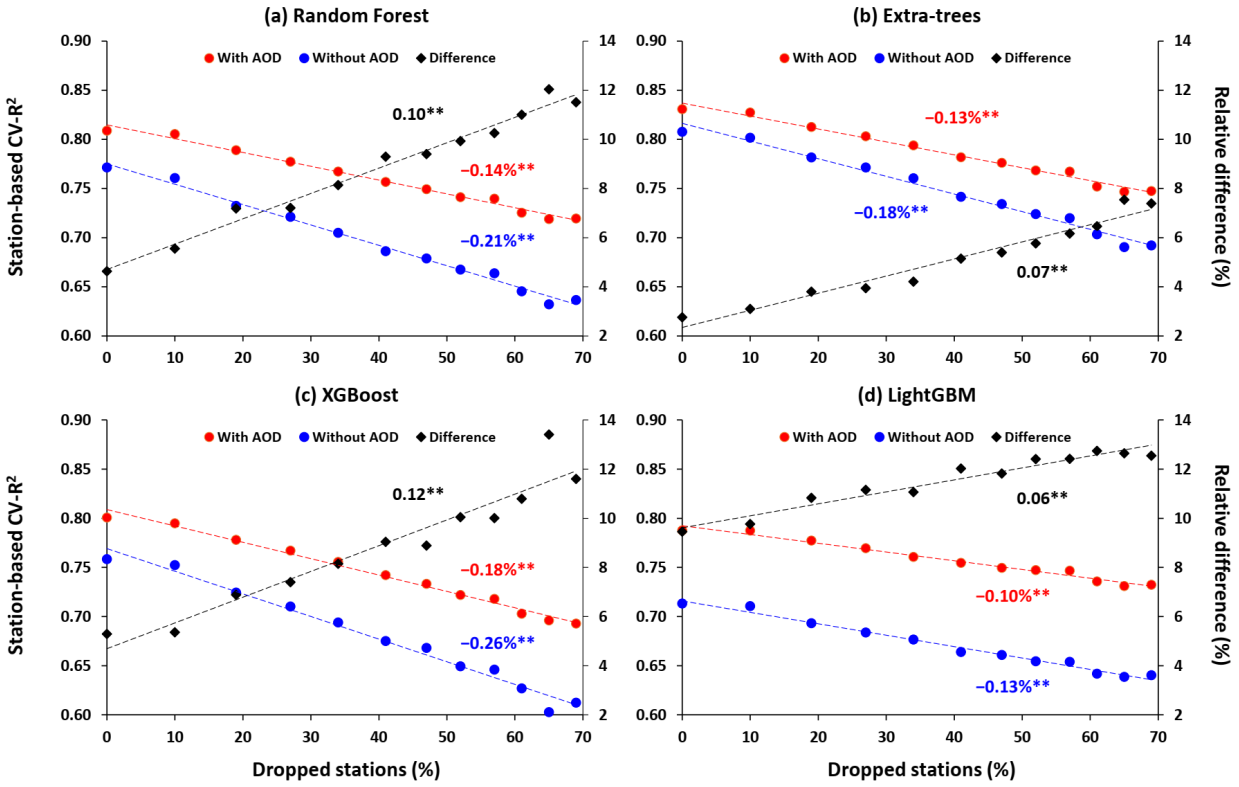

Figure 6. Variation in the predictive ability (station-based CV-R^2, left ordinate) and relative difference (%, right ordinate) of daily $PM_{2.5}$ estimates as a function of decreasing station density for (**a**) Random Forest, (**b**) Extra-trees, (**c**) XGBoost, and (**d**) LightGBM, with (red dots) and without (blue dots) including satellite AOD as an input predictor. Slopes of the best-fit lines from linear regression for each set of results are given, where ** denotes the 99% confidence ($p < 0.01$) level. Black dashed lines represent the regressed fits of the relative difference between each set of results.

The contrasting results with comparable superior accuracy in $PM_{2.5}$ estimation without incorporating AOD input can be attributed to the high density of ground observation stations in the specific study area. In areas with a sufficiently dense network of monitoring sites, excluding AOD can still lead to relatively accurate $PM_{2.5}$ estimations, which largely benefit from the presence of spatial autocorrelation in air pollution, e.g., $PM_{2.5}$ levels and auxiliary factors such as meteorological fields are highly similar in neighboring locations. As a result, $PM_{2.5}$ concentrations in nearby sites can be reasonably estimated based on the established relationships between $PM_{2.5}$ and non-AOD factors from nearby stations. Nevertheless, as the station density decreases, the spatial autocorrelation weakens, and the disparity between natural and human-influenced conditions grows, and consequently, the prediction error rapidly increases. This highlights the critical role of AOD in areas with limited ground monitoring because it significantly enhances the accuracy of $PM_{2.5}$ predictions by providing crucial background pollution information, compensating for the lack of ground observation data.

4. Conclusions

Machine learning (ML) has been used widely to infer ground-level $PM_{2.5}$ using satellite-retrieved aerosol optical depth (AOD) to fill large gaps between $PM_{2.5}$ stations without quantification of its contribution, which is the objective of this study. We rigorously and quantitatively assess the contribution of AOD to the ML-based estimation of $PM_{2.5}$ by applying four common ML models (the Random Forest model, the Extra-trees model, the XGBoost model, and the LightGBM model) to ample measurements from China's high-density $PM_{2.5}$ observation network, the MODIS Multi-Angle Implementation of Atmospheric Correction satellite AOD retrieval product, and many other ancillary meteorological and environmental data from the eastern half of China. Two assessment methods are used, i.e., feature importance (FI) and permutation importance (PI). The contribution of AOD is also assessed by comparing the retrieval results obtained by including and not including AOD. All assessment tests are made for varying numbers of $PM_{2.5}$ stations whose data are sampled by station-based and sample-based 10-CV.

The major findings are summarized as follows: (1) As the station density decreases, the FI and PI of AOD in the four ML models have clear upward trends. This trend indicates the importance and contribution of AOD to improving the accuracy of estimating $PM_{2.5}$, becoming more pronounced in areas with sparse observation stations. (2) As the density of observation stations decreases, the ML models without AOD exhibit a more pronounced decline in overall accuracy compared to the models that incorporate AOD. Additionally, for every 10% reduction in the number of stations, the uncertainty in the estimated accuracy increases by approximately 0.7–1.2%. (3) As the station density decreases, the ML models without AOD exhibit a faster decline in predictive ability compared with these models with AOD. On average, for every 10% reduction in the number of stations, the uncertainty in the predicted accuracy increases by approximately 0.6–1.2%. These findings demonstrate the indispensable role of AOD in any ML model to effectively counteract the negative impact of no or sparse $PM_{2.5}$ stations, resulting in improved accuracy for both estimating and predicting $PM_{2.5}$ levels.

AOD represents the degree of light attenuation caused by the scattering and absorption of atmospheric aerosols in the vertical direction and has served as a crucial indicator in deriving surface particulate matter concentrations. The importance of AOD was confirmed through a sensitivity analysis showing that satellite AOD has the highest FI and PI values in $PM_{2.5}$ modeling using various ML models. As the number of ground stations decreases, the AOD contribution is more apparent, as is the faster drop in ML model performance without AOD. This further underlines how using satellite AOD is essential, providing key background pollution information in areas without ground stations, thereby improving the prediction capability of ground-based $PM_{2.5}$. On the contrary, in this case, relying solely on auxiliary factors such as meteorological fields is far from adequate.

Even though the chosen area in eastern China benefits from a dense and reasonably evenly distributed network of ground observation stations for $PM_{2.5}$, enhancing its representativeness, there still exists the question of the uniformity of the spatial distribution of stations. Further analysis incorporating spatial-block cross-validation is needed to effectively reduce the impact of this issue. This will be undertaken in our future study. Additionally, this approach could be applied to PM_{10}, considering its high similarities with $PM_{2.5}$, i.e., AOD retains its significance as a crucial input variable in PM_{10} predictions [58]. However, additional sensitivity analyses are warranted to confirm this hypothesis. Regarding other pollutants, since they possess entirely different key input variables, further investigations are needed to accurately understand their behaviors [e.g., the importance of satellite tropospheric NO_2 in surface NO_2 modelling [59]].

Author Contributions: Conceptualization: Z.L. and J.W.; Methodology: J.W., Z.T. and Z.L.; Analysis: Z.T. and J.W.; Writing: Z.T., J.W. and Z.L.; Funding Acquisition: Z.L. All authors have read and agreed to the published version of the manuscript. Zhanqing Li's contribution to this publication was not part of his University of Maryland duties or responsibilities.

Funding: This research was funded by the National Natural Science Foundation (42030606).

Data Availability Statement: The $PM_{2.5}$ observation station data used in this study are real-time ground-measured air quality data (including $PM_{2.5}$) in China from the China National Environment Monitoring Center (http://www.cnemc.cn, accessed on 1 January 2023). The NASA data center (https://ladsweb.modaps.eosdis.nasa.gov/search/, accessed on 1 January 2023) provides MODIS MAIAC aerosol products. Meteorological reanalysis data are collected from the fifth-generation European Reanalysis Interim dataset (ERA5, https://www.ecmwf.int/, accessed on 1 January 2023) released by the European Centre for Medium-Range Weather Forecasts.

Acknowledgments: We thank M. Cribb from the University of Maryland for helping in editing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IPCC 2021. *Climate Change, 2021: The Physical Science Basis*; Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change IPCC Working Group I Contribution to AR5Rep.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2021.
2. Guo, J.; Xia, F.; Zhang, Y.; Liu, H.; Li, J.; Lou, M.; He, J.; Yan, Y.; Wang, F.; Min, M.; et al. Impact of diurnal variability and meteorological factors on the $PM_{2.5}$-AOD relationship: Implications for $PM_{2.5}$ remote sensing. *Environ. Poll.* **2017**, *221*, 94. [CrossRef] [PubMed]
3. Li, Z.; Wang, Y.; Guo, J.; Zhao, C.; Cribb, M.C.; Dong, X.; Fan, J.; Gong, D.; Huang, J.; Jiang, M.; et al. East Asian Study of Tropospheric Aerosols and their Impact on Regional Clouds, Precipitation, and Climate (EAST-AIR_{CPC}). *J. Geophys. Res. Atmos.* **2019**, *124*, 13026–13054. [CrossRef]
4. Duyzer, J.; van den Hout, D.; Zandveld, P.; van Ratingen, S. Representativeness of air quality monitoring networks. *Atmos. Environ.* **2015**, *104*, 88–101. [CrossRef]
5. Alsahli, M.M.; Al-Harbi, M. Allocating optimum sites for air quality monitoring stations using GIS suitability analysis. *Urban Clim.* **2018**, *24*, 875–886. [CrossRef]
6. Chen, N.; Yang, M.; Du, W.; Min, H. $PM_{2.5}$ estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution MAIAC AOD in Hubei, China. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 31. [CrossRef]
7. van Donkelaar, A.; Martin, R.V.; Park, R.J. Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing. *J. Geophys. Res. Atmos.* **2006**, *111*, D21201. [CrossRef]
8. Ma, X.; Wang, J.; Yu, F.; Jia, H.; Hu, Y. Can MODIS AOD be employed to derive $PM_{2.5}$ in Beijing-Tianjin-Hebei over China? *Atmos. Res.* **2016**, *181*, 250–256. [CrossRef]
9. Li, S.; Joseph, E.; Min, Q. Remote sensing of ground-level $PM_{2.5}$ combining AOD and backscattering profile. *Remote Sens. Environ.* **2016**, *183*, 120–128. [CrossRef]
10. Li, Z.; Goloub, P.; Devaux, C.; Gu, X.; Qiao, Y.; Zhao, F.; Chen, H. Aerosol polarized phase function and single-scattering albedo retrieved from ground-based measurements. *Atmos. Res.* **2004**, *71*, 233–241. [CrossRef]
11. Gupta, P.; Christopher, S.A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40*, 5880–5892. [CrossRef]
12. Kumar, N. What can affect AOD–$PM_{2.5}$ association? *Environ. Health Perspect.* **2010**, *118*, A109–A110. [CrossRef] [PubMed]

13. Wang, J.; Christopher, S.A. Intercomparison between satellite-derived aerosol optical thickness and $PM_{2.5}$ mass: Implications for air quality studies. *Geophys. Res. Lett.* **2003**, *30*, 2095. [CrossRef]
14. Natunen, A.; Arola, A.; Mielonen, T.; Huttunen, J.; Lehtinen, K.E.J. A multi-year comparison of $PM_{2.5}$ and AOD for the Helsinki region. *Boreal Environ. Res.* **2010**, *15*, 544–552. [CrossRef]
15. Kloog, I.; Nordio, F.; Coull, B.A.; Schwartz, J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the Mid-Atlantic states. *Environ. Sci. Technol.* **2012**, *46*, 11913–11921. [CrossRef] [PubMed]
16. Ma, Z.; Hu, X.; Sayer, A.M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-based spatiotem-poral trends in $PM_{2.5}$ concentrations: China 2004-2013. *Environ. Health Perspect.* **2016**, *124*, 184–192. [CrossRef]
17. Qu, W.; Wang, J.; Zhang, X.; Sheng, L.; Wang, W. Opposite seasonality of the aerosol optical depth and the surface particulate matter concentration over the North China Plain. *Atmos. Environ.* **2016**, *127*, 90–99. [CrossRef]
18. Su, T.; Li, Z.; Kahn, R. Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: Regional pattern and influencing factors. *Atmos. Chem. Phys.* **2018**, *18*, 15921–15935. [CrossRef]
19. Koelemeijer, R.B.A.; Homan, C.D.; Matthijsen, J. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmos. Environ.* **2006**, *40*, 5304–5315. [CrossRef]
20. Brauer, M.; Amann, M.; Burnett, R.T.; Cohen, A.; Dentener, F.; Ezzati, M.; Henderson, S.B.; Krzyzanowski, M.; Martin, R.V.; Van Dingenen, R.; et al. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.* **2012**, *46*, 652–660. [CrossRef]
21. Chen, J.; de Hoogh, K.; Gulliver, J.; Hoffmann, B.; Hertel, O.; Ketzel, M.; Bauwelinck, M.; van Donkelaar, A.; Hvidtfeldt, U.A.; Katsouyanni, K.; et al. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environ. Int.* **2019**, *130*, 104934. [CrossRef]
22. Gupta, P.; Christopher, S. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res. Atmos.* **2009**, *114*, D14205. [CrossRef]
23. Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. *Environ. Sci. Tech.* **2014**, *48*, 7436–7444. [CrossRef] [PubMed]
24. You, W.; Zang, Z.; Zhang, L.; Li, Y.; Wang, W. Estimating national-scale ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on MODIS and MISR AOD. *Environ. Sci. Pollut. Res.* **2016**, *23*, 8327–8338. [CrossRef]
25. He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. *Remote Sens. Environ.* **2018**, *206*, 72–83. [CrossRef]
26. Xiao, Q.; Wang, Y.; Chang, H.H.; Meng, X.; Liu, Y. Full-coverage high-resolution daily $PM_{2.5}$ estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sens. Environ.* **2017**, *199*, 437–446. [CrossRef]
27. Liu, Y.; Sarnat, J.A.; Kilaru, A.; Jacob, D.J.; Koutrakis, P. Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing. *Environ. Sci. Technol.* **2005**, *39*, 3269–3278. [CrossRef]
28. Lee, H.J.; Liu, Y.; Coull, B.A.; Schwartz, J.; Koutrakis, P. A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations. *Atmos. Chem. Phys.* **2011**, *11*, 7991–8002. [CrossRef]
29. Chen, W.; Ran, H.; Cao, X.; Wang, J.; Zheng, X. Estimating $PM_{2.5}$ with high-resolution 1-km AOD data and an improved machine learning model over Shenzhen, China. *Sci. Total Environ.* **2020**, *746*, 141093. [CrossRef]
30. Wei, J.; Huang, W.; Li, Z.; Xue, W.; Cribb, M. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **2019**, *231*, 111221. [CrossRef]
31. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1-km-resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. [CrossRef]
32. Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. *Remote Sens. Environ.* **2021**, *252*, 112136. [CrossRef]
33. Pan, B. Application of XGBoost algorithm in hourly $PM_{2.5}$ concentration prediction. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *113*, 012127. [CrossRef]
34. Wei, J.; Li, Z.; Pinker, R.T.; Sun, L.; Li, R. Himawari-8-derived diurnal variations of ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). *Atmos. Chem. Phys.* **2021**, *21*, 7863–7880. [CrossRef]
35. Fang, X.; Zou, B.; Liu, X.; Sternberg, T.; Zhai, L. Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling. *Remote Sens. Environ.* **2016**, *186*, 152–163. [CrossRef]
36. Meng, X.; Fu, Q.; Ma, Z.; Chen, L.; Zou, B.; Zhang, Y.; Xue, W.; Wang, J.; Wang, D.; Han, H. Estimating ground-level PM_{10} in a Chinese city by combining satellite data, meteorological information and a land use regression model. *Environ. Pollut.* **2016**, *208*, 177–184. [CrossRef] [PubMed]
37. Pereira, G.; Lee, H.J.; Bell, M.; Regan, A.; Malacova, E.; Mullins, B.; Knibbs, L.D. Development of a model for particulate matter pollution in Australia with implications for other satellite-based models. *Environ. Res.* **2017**, *159*, 9–15. [CrossRef] [PubMed]
38. Chen, G.; Li, Y.; Zhou, Y.; Shi, C.; Liu, Y. The comparison of AOD-based and non-AOD prediction models for daily $PM_{2.5}$ estimation in Guangdong province, China with poor AOD coverage. *Environ. Res.* **2021**, *195*, 110735. [CrossRef] [PubMed]

39. Yu, W.; Li, S.; Ye, T.; Xu, R.; Song, J.; Guo, Y. Deep ensemble machine learning framework for the estimation of $PM_{2.5}$ concentrations. *Environ. Health Perspect.* **2022**, *130*, 037004. [CrossRef]
40. Lyapustin, A.; Martonchik, J.; Wang, Y.; Laszlo, I.; Korkin, S. Multi-Angle Implementation of Atmospheric Correction (MAIAC): 1. Radiative transfer basis and look-up tables. *J. Geophys. Res. Atmos.* **2011**, *116*, D03210. [CrossRef]
41. Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765. [CrossRef]
42. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
43. Peuch, V.H.; Engelen, R.; Ades, M.; Barre, J.; Suttie, M. The use of satellite data in the Copernicus Atmosphere Monitoring Service (CAMS). In *IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Manhattan, NY, USA, 2018.
44. Wei, J.; Li, Z.; Wang, J.; Li, C.; Gupta, P.; Cribb, M. Ground-level gaseous pollutants (NO_2, SO_2, and CO) in China: Daily seamless mapping and spatiotemporal variations. *Atmos. Chem. Phys.* **2023**, *23*, 1511–1532. [CrossRef]
45. Malakar, N.K.; Lary, D.J.; Moore, A.; Gencaga, D.; Roscoe, B.; Albayrak, A.; Petrenko, M.; Wei, J. Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing. In Proceedings of the 2012 Conference on Intelligent Data Understanding (CIDU 2012), Boulder, CO, USA, 24–26 October 2012.
46. Lary, D.J.; Faruque, F.S.; Malakar, N.; Moore, A.; Roscoe, B.; Adams, Z.L.; Eggelston, Y. Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$). *Geospat. Health* **2014**, *8*, S611–S630. [CrossRef]
47. Reid, C.E.; Jerrett, M.; Petersen, M.L.; Pfister, G.G.; Morefield, P.E.; Tager, I.B.; Raffuse, S.M.; Balmes, J.R. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ. Sci. Technol.* **2015**, *49*, 3887–3896. [CrossRef] [PubMed]
48. Breiman, L. Random forests. *Mach. Learn.* **2002**, *45*, 5–32. [CrossRef]
49. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.S.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. [CrossRef]
50. Hu, X.; Belle, J.H.; Meng, X.; Wildani, A.; Waller, L.A.; Strickland, M.J.; Liu, Y. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Tech.* **2017**, *51*, 6936–6944. [CrossRef]
51. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
52. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
53. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A highly effificient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; ACM: Long Beach, CA, USA, 2017; pp. 3149–3157. Available online: https://dl.acm.org/doi/10.5555/3294996.3295074 (accessed on 1 January 2020).
54. Loecher, M. Unbiased variable importance for random forests. *Commun. Stat. Theory Methods* **2022**, *51*, 1413–1425. [CrossRef]
55. Kim, H.; Loh, W.-Y. Classification trees with unbiased multiway splits. *J. Am. Stat. Assoc.* **2001**, *96*, 589–604. [CrossRef]
56. Rodriguez, J.; Perez, A.; Lozano, J. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [CrossRef] [PubMed]
57. Wei, J.; Li, Z.; Chen, X.; Li, C.; Sun, Y.; Wang, J.; Lyapustin, A.; Brasseur, G.; Jiang, M.; Sun, L.; et al. Separating daily 1-km $PM_{2.5}$ inorganic chemical composition in China since 2000 via deep learning integrating ground, satellite, and model data. *Environ. Sci. Tech.* **2023**. [CrossRef] [PubMed]
58. Wei, J.; Li, Z.; Xue, W.; Sun, L.; Fan, T.; Liu, L.; Su, T.; Cribb, M. The ChinaHighPM_{10} dataset: Generation, validation, and spatiotemporal variations from 2015 to 2019 across China. *Environ. Int.* **2021**, *146*, 106290. [CrossRef] [PubMed]
59. Wei, J.; Liu, S.; Li, Z.; Liu, C.; Qin, K.; Liu, X.; Pinker, R.; Dickerson, R.; Lin, J.; Boersma, K.; et al. Ground-level NO_2 surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence. *Environ. Sci. Tech.* **2022**, *56*, 9988–9998. [CrossRef] [PubMed]