applied sciences

MDPI

Article

Evaluation of Machine Learning Models for Estimating $PM_{2.5}$ Concentrations across Malaysia

Nurul Amalin Fatihah Kamarul Zaman [1], Kasturi Devi Kanniah [1,2,*], Dimitris G. Kaskaoutis [3,4,*] and Mohd Talib Latif [5]

1 Tropical Map Research Group, Faculty of Built Environment & Surveying, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia; nafatihah4@graduate.utm.my
2 Centre for Environmental Sustainability and Water Security (IPASA), Research Institute for Sustainable Environment, Universiti Teknologi Malaysia, Skudai 81310 UTM, Johor, Malaysia
3 Institute for Environmental Research and Sustainable Development, National Observatory of Athens, 15236 Athens, Greece
4 Environmental Chemical Processes Laboratory, Department of Chemistry, University of Crete, 71003 Crete, Greece
5 Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; talib@ukm.edu.my
* Correspondence: kasturi@utm.my (K.D.K.); dkask@noa.gr (D.G.K.)

Abstract: Southeast Asia (SEA) is a hotspot region for atmospheric pollution and haze conditions, due to extensive forest, agricultural and peat fires. This study aims to estimate the $PM_{2.5}$ concentrations across Malaysia using machine-learning (ML) models like Random Forest (RF) and Support Vector Regression (SVR), based on satellite AOD (aerosol optical depth) observations, ground measured air pollutants (NO_2, SO_2, CO, O_3) and meteorological parameters (air temperature, relative humidity, wind speed and direction). The estimated $PM_{2.5}$ concentrations for a two-year period (2018–2019) are evaluated against measurements performed at 65 air-quality monitoring stations located at urban, industrial, suburban and rural sites. $PM_{2.5}$ concentrations varied widely between the stations, with higher values (mean of 24.2 ± 21.6 μg m^{-3}) at urban/industrial stations and lower (mean of 21.3 ± 18.4 μg m^{-3}) at suburban/rural sites. Furthermore, pronounced seasonal variability in $PM_{2.5}$ is recorded across Malaysia, with highest concentrations during the dry season (June–September). Seven models were developed for $PM_{2.5}$ predictions, i.e., separately for urban/industrial and suburban/rural sites, for the four dominant seasons (dry, wet and two inter-monsoon), and an overall model, which displayed accuracies in the order of R^2 = 0.46–0.76. The validation analysis reveals that the RF model (R^2 = 0.53–0.76) exhibits slightly better performance than SVR, except for the overall model. This is the first study conducted in Malaysia for $PM_{2.5}$ estimations at a national scale combining satellite aerosol retrievals with ground-based pollutants, meteorological factors and ML techniques. The satisfactory prediction of $PM_{2.5}$ concentrations across Malaysia allows a continuous monitoring of the pollution levels at remote areas with absence of measurement networks.

Keywords: $PM_{2.5}$; Himawari-8; random forest; support vector regression; air pollution; Malaysia

Citation: Zaman, N.A.F.K.; Kanniah, K.D.; Kaskaoutis, D.G.; Latif, M.T. Evaluation of Machine Learning Models for Estimating $PM_{2.5}$ Concentrations across Malaysia. *Appl. Sci.* **2021**, *11*, 7326. https://doi.org/10.3390/app11167326

Academic Editor: Yves Rybarczyk

Received: 25 June 2021
Accepted: 5 August 2021
Published: 9 August 2021

1. Introduction

Air pollution has become an acute environmental and health issue in developing countries during the last decades due to intense industrialization and urbanization processes [1–3]. It is estimated that about 7 million people die every year worldwide because of exposure to fine particulate pollution < 2.5 μm ($PM_{2.5}$), while about 91% of the world's population live in areas with $PM_{2.5}$ concentrations above the allowable limits of 10–20 μg m^{-3} [4]. Southeast Asia (SEA) is not an exception to high pollution levels and experiences persistent haze conditions, especially during the dry season (June to September) due to extensive forest, agricultural and peat fires [5–9]. Malaysia is located in the main pathway of the SEA pollution

outflow [10] that escalates the pollution levels due to trans-boundary aerosol transport. Air quality in Malaysia is considered rather degraded, as the annual mean $PM_{2.5}$ concentration is about 20 μg m^{-3} in 2019 [11], thus exceeding the limit set by WHO. Local sources such as traffic and industrial emissions, as well as biomass burning significantly contribute to the local pollution and aerosol loading [12–14]. According to the Department of Statistics Malaysia [15], chronic lower respiratory diseases (CLRD) is the fifth leading cause of death in Malaysia that was increased from 2.0% to 2.6% between the years 2017 and 2018. In fact, chronic exposure to $PM_{2.5}$ has significant impacts on human health by causing asthma, chronic obstructive pulmonary disease, lung cancer, cardiovascular and neurotoxic effects [16,17]. Therefore, continuous monitoring of the levels and mapping of spatial distribution of $PM_{2.5}$ is especially important for taking appropriate actions to maintain a good air quality over Malaysia [18,19].

Nowadays, many satellites provide reliable AOD (aerosol optical depth) products that can be used through various techniques for PM estimations from space, such as the Advanced Very High Resolution Radiometer (AVHRR) [20,21], Multiangle Imaging Spectroradiometer (MISR) [22], Medium Resolution Imaging Spectrometer (MERIS) [23], Spinning Enhanced Visible and Infrared Imager (SEVIRI) [24] and Moderate Resolution Imaging Spectroradiometer (MODIS) [25–28]. Among the satellite data products, AOD provided by MODIS sensor on board Terra and Aqua satellites has been widely used for many applications due to its high retrieval accuracy over land and near daily global coverage at 10 km and 3 km spatial resolution [26,27,29]. However, its usability in SEA is still limited due to large missing data series as a result of heavy and extended cloudiness [18,30–32]. On the other hand, geostationary satellites provide high temporal resolution (~15 min) data, thus limiting the problem of cloudiness. In the past few years, geostationary satellites such as Geostationary Operational Environmental Satellite (GOES) [33], Geostationary Ocean Color Imager (GOCI) [34], Fengyun-4 [35] and Himawari-8 [36–38] are available and provide continuous data over the SEA region.

Significant progress has been made in developing and establishing various techniques for estimating PM concentrations from space at local, regional and global scales. The progress in usage of linear and multi-linear statistics, regression-based, machine learning and hybrid models for estimations of $PM_{2.5}$ and PM_{10} concentrations during the last decades is reviewed in recent works [1,39,40]. In the early 2000s, most researchers predicted PM using only AOD, by means of simple linear regression techniques [41,42]. Later on, more advanced techniques were developed to incorporate AOD and other important parameters that may influence PM distribution spatially and temporarily, starting from multiple linear regressions [18,43–46], chemical transport models (CTM) [47–50], mixed effect models (MEM) [51–54], artificial neural networks (ANN) [18,55–58], geographic weighted regression (GWR) [59–62] and generalized additive models (GAM) [63–66]. These techniques were used to capture the non-linear relationships that exist between the variables. Consequently, complex techniques have been developed by combining two or more statistical techniques; for instance, merging MEM and GWR [67] or incorporating MEM into GAM [68]. Nowadays, Machine Learning (ML) techniques such as deep neural network (DNN), support vector regression (SVR) and random forest (RF) enable to capture the complex relationships between parameters, exhibiting greater performance in estimating $PM_{2.5}$ [69,70] and are increasingly used in air quality studies [71–74]. Furthermore, some studies have incorporated meteorological factors with land use variables to predict the spatial and temporal variation of aeolian erosion and PM [75–79]. However, only few studies have incorporated air pollutant concentrations for $PM_{2.5}$ estimations [65,80,81]. Song et al. [65] explored the use of generalized additive model to estimate $PM_{2.5}$ concentrations in the Xi'an City, China (3581 square kilometres) using a combination of air pollutants (SO_2, CO, NO_2, and O_3), AOD and meteorological variables. The model was found to explain ~70% of the variance in $PM_{2.5}$ concentrations, with CO concentration and AOD represented most of the variation. The influence of air pollutants on the seasonal variability of $PM_{2.5}$ in an urban-industrial environment in Malaysia was investigated by [81] for one year. This study concluded that only gases (CO, NO_2, NO and SO_2) significantly affected

the $PM_{2.5}$ mass, but not the meteorological factors (rainfall, wind speed and wind direction). Based on the prediction model of [80], temperature, Normalized Difference Vegetation Index (NDVI), humidity and residential area were found to be important parameters affecting the spatial variation of $PM_{2.5}$ in Jakarta, Indonesia. The same study [80], also showed that several parameters (PM_{10}, NO_2, SO_2, UV, rainfall, land use and NDVI) influenced the distribution of $PM_{2.5}$ in Taipei, Taiwan. More studies are therefore needed to find out the role of gases and meteorology in affecting the spatial and seasonal patterns of $PM_{2.5}$, even using meteorological normalization techniques in order to exclude the effect of changing meteorology on PM concentrations trends [82]. Recently, ML approaches (random forest regression models) were implemented to predict the large reductions in air pollutants, i.e., PM_{10}, NO_2, O_3, during the COVID-19 lockdown period [83]. Furthermore, PM_{10}, NO_2 and carbonaceous aerosols (organic carbon, elemental carbon) were also used in ML techniques (Lasso, Random Forest, AdaBoost, Support Vector Machine and Partials Least squares) for analysing air pollutants at street canyons [84]. Studies dealing with PM estimations in Malaysia are rather limited [18,30]. Shaziayani et al. [85] has reviewed PM_{10} modelling studies in Malaysia, and only four studies used ML techniques in predicting PM_{10}. On the other hand, $PM_{2.5}$ studies are even fewer and most of them in Malaysia have been performed at small spatial scales [86,87].

The current study is the first in Malaysia and one of the very few works conducted worldwide aiming to estimate the $PM_{2.5}$ concentrations at a large (national) scale using pollution gases, AOD and meteorological factors based on machine-learning techniques. In order to extend the spatial coverage to the whole country (both Peninsular and Island Malaysia) and aiming to improve the accuracy of $PM_{2.5}$ estimates, this study integrates hourly AOD products from Himawari-8 satellite sensor, along with meteorological parameters and gaseous pollutants using machine learning techniques, i.e., random forest (RF) and Support Vector Regression (SVR). The models were developed separately for urban/industrial, suburban/rural sites and for the four dominant seasons in order to better represent the spatial (between sites) and temporal (between seasons) variation of the $PM_{2.5}$ concentrations. Variable importance analysis was conducted to identify the primary parameters that affect the $PM_{2.5}$ concentrations in Malaysia in a way to develop regression models for estimations of $PM_{2.5}$. The performance of RF and SVR was evaluated at different seasons and locations against measured $PM_{2.5}$ concentrations. The results will assist in representing the spatial and temporal evolution of $PM_{2.5}$ in Malaysia and for establishing measures in a way to improve air quality across the country.

Section 2 briefly describes the study area; Section 3 refers to the dataset that was used as variables in the prediction models; Section 4 refers to the $PM_{2.5}$ measurements across Malaysia. The results and model evaluation are included in Section 5, while Section 6 summarises the conclusions.

2. Study Area

Malaysia is one of the developed countries in SEA region with a rapid urbanization rate since 1970 [88]. Consequently, air pollution has become one of the serious environmental and human health concerns across the country [5,89], particularly in urban, industrialised and congested traffic areas such as Klang Valley (in the west coast of Peninsular Malaysia), Johor Bahru (southern tip of the Peninsula) and Georgetown, Penang (north of Peninsula) [90,91]. Air quality deteriorates at several parts of Peninsular Malaysia and in Borneo Island during the dry season mainly due to trans-boundary haze from neighbouring countries and regional/local forest fires. The concentrations of aerosols and air pollutants display a distinct seasonality, influenced by local meteorological conditions, i.e., rainfall, wind speed, relative humidity (RH) and temperature [86], being lower during the monsoon rainy season (November–March). In this study, 65 air quality monitoring stations (Figure 1) distributed across the Peninsular and Island Malaysia (Labuan, Sabah and Sarawak) were used to analyse the air pollution levels ($PM_{2.5}$, SO_2, NO_2, CO and O_3

concentrations). The stations are representative of industrial (7 stations), urban (10 stations), suburban (36 stations) and rural (12 stations) areas.

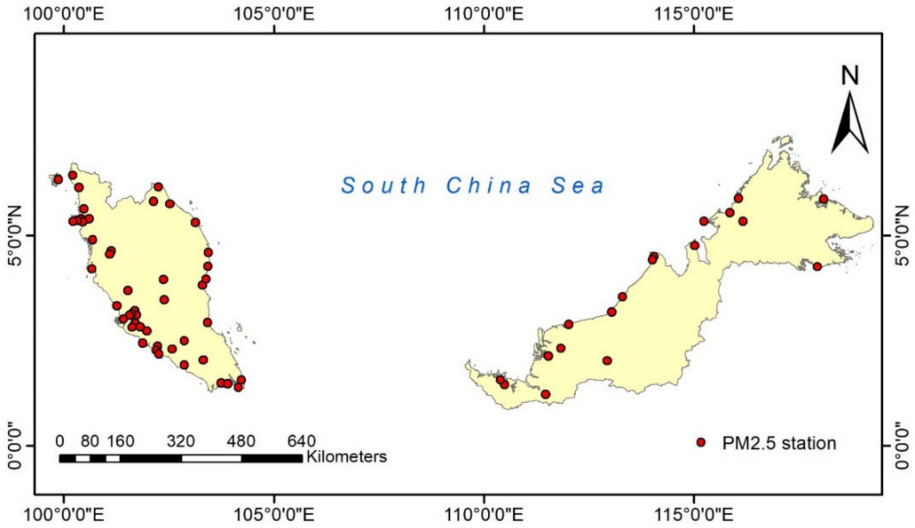

Figure 1. Locations of the 65 air quality monitoring stations across Malaysia.

3. Dataset

The dataset used in this study consist of $PM_{2.5}$ and air pollutant concentrations, along with meteorological parameters from ground stations, as well as AOD data from Himawari-8 satellite.

3.1. Ground Measurements

$PM_{2.5}$ concentrations in Malaysia have been measured since April 2017. The Department of Environment, Malaysia (DOE) increased the number of air quality monitoring stations across the country from 52 to 65 stations in 2017 (Figure 1). Furthermore, these stations also measure meteorological parameters (e.g., ambient temperature, TEMP; RH, wind speed, WS; wind direction, WD) and gaseous pollutants (e.g., nitrogen dioxide, NO_2; carbon monoxide, CO; sulphur dioxide, SO_2; ozone, O_3). The stations are strategically distributed to represent urban, industrial, suburban and rural areas [92]. $PM_{2.5}$ measurements were performed via the TEOM 1405DF, which is a continuous dichotomous ambient air monitoring system with two Filter Dynamics Measurements Systems [93], able to measure $PM_{2.5}$ and PM_{10}. SO_2, NO_2, CO and O_3 are measured using Thermo Scientific model 43i, model 42i, model 48i and model 49i, respectively [93,94]. RH and TEMP were recorded using a Climatronic AIO 2 Weather Sensor (Climatronic Corporation) [95]. All the ground data were obtained on an hourly basis covering the period from January 2018 to December 2019.

All air quality and meteorological measurements went through quality assurance and quality control (QA/QC) procedures. Instruments for the detection of gases were manually calibrated once a fortnight. Flow verification for PM_{10} and $PM_{2.5}$ measurements using TEOM was conducted once a month. The data removal during the QC check was predominantly due to insufficient measurements and instrument failure, while some perturbed data were also excluded as outliers in a second-level QC check [95].

3.2. Satellite Data

Himawari-8 is a geostationary satellite operated by the Japan Meteorological Agency. It was launched on 7 October 2014 and carries the Advanced Himawari Imager (AHI) sensor, which is equipped with 16 bands from visible to infrared [36]. Himawari-8 releases AOD products at two levels, namely Level 2 (10 min temporal) and Level 3 (hourly and daily), which have been used for various applications including estimation of PM [96–99],

dust detection [100] and aerosol data assimilation [101]. The L3 product is an improved version of the L2 AOD product that minimized cloud contamination [102] and has a 5 km spatial resolution. Himawari-8 AOD at 500 nm is associated with quality assurance levels namely "very good", "good", "marginal" and "no confident (or no retrieval)" [99]. In this study, only the "very good" L3 AOD_{500} retrievals were considered for $PM_{2.5}$ estimations, downloaded from the Japan Aerospace Exploration Agency (JAXA) website: available online: http://www.eorc.jaxa.jp/ptree/index.html (accessed on 10 May 2021) for the period January 2018–December 2019. Recently, Himawari-8 AODs were used to estimate the $PM_{2.5}$ concentrations over Hubei province, China [103]. Application of the Himawari-8 L2 AOD data over Malaysia revealed an overestimation by 24.2% [104], while the L3 AOD products displayed a better agreement with the Aerosol Robotic Network (AERONET) AODs with a coefficient of determination R^2 = 0.81, root mean square error (*RMSE*) of 0.13 and an overall overestimation of only 1% [92]. Moreover, Himawari-8 AODs presented a good agreement with AERONET AODs in China (R^2 = 0.41–0.83; *RMSE* = 0.18–0.31) [105], Southeast Asia (R^2 = 0.64; *RMSE* = 0.28), East Asia (R^2 = 0.83; *RMSE* = 0.14) [106], Korea (R^2 = 0.69; *RMSE* = 0.19) and Beijing-Tianjin-Hebei (R^2 = 0.76; *RMSE* = 0.36) [107].

4. $PM_{2.5}$ Estimation

The overall methodology used for the $PM_{2.5}$ estimations over Malaysia is illustrated in Figure 2. The model inputs consist of hourly AOD, SO_2, NO_2, CO, O_3, WS, WD, TEMP and RH values. Hourly AOD data from Himawari-8 were extracted at 5 × 5 km over the air quality monitoring stations and temporally collocated with ground measurements. Wind direction was used in the model because wind blowing from a highly polluted area can influence air quality in other downwind places. Wind speed enables to accelerate pollutants travelling from other places but also contributes to the dilution processes at local level [108]. On the other hand, temperature can trigger biogenic emissions, photochemical reactions and secondary aerosol formation over the region [109] and also control the temperature inversions, which can trap the pollutants near the surface [45,108,110]. Finally, RH may affect the hygroscopic growth of particles and enhance the aerosol scattering [111–113].

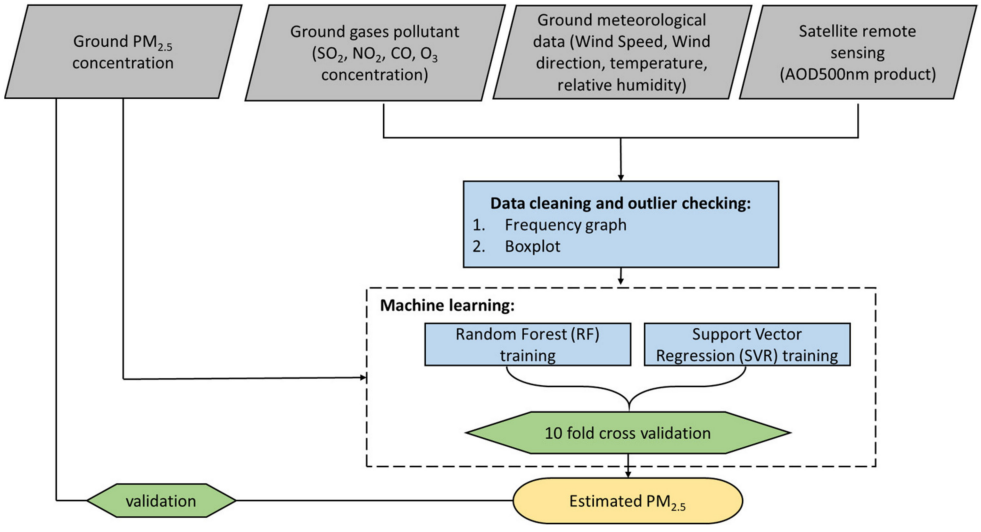

Figure 2. Flowchart of the overall dataset and methodology used for the estimations of $PM_{2.5}$ concentrations.

In this study, we utilized and evaluated two ML models, namely, Support Vector Regression (SVR) and Random Forest (RF) to estimate $PM_{2.5}$ concentrations at the 65 air quality monitoring stations across Malaysia. The input variables to the SVR and RF models are selected based on our previous study [18,31] and other literature [114,115]. In total, we developed 7 different models to represent the spatial and seasonal effects on $PM_{2.5}$ distributions. Model 1 considers all data from the 65 stations, but other models, i.e., models

2 and 3, only represent urban/industrial and suburban sites, respectively, while models 4 to 7 represent different seasons (wet, dry and two inter monsoon). The models are described in the subsequent sections.

4.1. Machine Learning (ML) for $PM_{2.5}$ Estimation

Nowadays, several studies have used machine learning (ML) techniques, including RF model, aiming to increase the accuracy in prediction of PM concentrations, since these models are flexible in nonlinear approaches [72,116–119]. The SVR and RF techniques were particularly selected in this study to achieve more accurate $PM_{2.5}$ estimations using satellite-derived AOD, meteorological parameters and gaseous pollutants as predictor variables. The Classification And REgression Training (CARET) package was used to perform the RF and SVR modelling. The data splitting, pre-processing, model tuning and variable importance analysis were executed in a R environment. SVR depends on the kernel function and due to its excellent generalization capability, it is able to minimize the overfitting [120], and therefore, it has been used for PM estimations [69,108,121]. SVR can fit the errors within a certain threshold by finding an appropriate boundary line (between hyperplane) to suit the data. The flexibility of SVR depends on the selection of the parameter such as kernel function, cost function and epsilon value. There are four types of kernel functions namely linear, polynomial, sigmoid and radial basis function (RBF) that were used in this study for capturing the non-linear dynamics [69], whilst cost function was used to avoid any overfitting of the data, as small cost value leads to large margin (or wide boundary line) and causes overfitting in the model. The epsilon value controls the number of support vectors used to develop the regression function, while the smaller epsilon value indicates an optimum accuracy. Initially, the SVR parameters are selected based on trial-and-error values, but we found that the default values (Supplementary Materials Table S1) included in the R package "e1071" provided the most promising results.

Random Forest (RF) is a tree-based ML technique proposed by [122]. Theoretically, RF model is an ensemble of multiple decision trees and uses the majority vote/decision of the trees as the final RF model [123,124]. The algorithm becomes more robust when more decision trees are constructed. RF randomly selects parameters in order to develop each tree, and therefore, it reflects potentially complex effects of predictors on the prediction [125]. The purpose of selecting random predictors instead of all predictors is to reduce the correlation between trees in order to make them disparate [126]. Thus, the variance of the RF prediction can reduce any overfittings. The number of decision trees can be modified to reduce the training time according to a required accuracy and computing capability [127]. In this study, RF model was run using the "Random Forest" in the R package. Since, RF is a non-parametric algorithm, here we only set the two most important parameters—although RF can have more parameters—which are mtry and ntree. Parameter mtry is a number of predictors sampled for splitting at each node while ntree is a number of trees in the forest. If mtry value is too small, it might be none of significant parameters included in the subset, and the insignificant parameters would be selected for a split. Therefore, the trees have poor predictive ability [126]. In this study, we set the mtry = 3 (as default: mtry = p/3, where p is the number of parameters used in the model), while ntree is set as 500 in the model. For tuning, we only tune mtry because the CARET package has automatic tuning for *mtry* only. Therefore, in this study, for ntree, we used the default value [128,129]. The results were obtained based on best mtry tuning accuracy. It should be mentioned that a limitation of this study is that models were not broadly optimized.

4.2. Model Validation

The total number of the matching dataset, covering all parameters at the 65 stations in Malaysia from 2018 to 2019, is 13,376. The matching data were randomly partitioned at a fraction of 70% for model calibration (model development) and 30% for model validation. In the model development, a sample based 10-fold cross validation technique was used, where the calibration data were randomly divided into 10 subsets; at any single moment,

one subset was used for validation and the remaining subsets were used for calibrating the model. The average value of the results of the 10 subsets was adopted as the model accuracy. The sample based 10 cross validation (CV) performed validation with matchup sample from both spatial and temporal dimension. This is a commonly used CV-based technique to reveal the overall predictive ability of $PM_{2.5}$ estimation models [130]. Then, the final models were validated using the 30% of the remaining data. Statistical indicators such as the coefficient of determination (R^2), Root Mean Square Error (*RMSE*), mean bias error (*MBE*) and Nash-Sutcliffe Efficiency (*NSE*) were used to evaluate the accuracy of the models. The *NSE* is a normalized statistic, which can determine the magnitude of the residual variance to the measured data variance and indicates how well the measured $PM_{2.5}$ versus estimated $PM_{2.5}$ data fits the 1:1 line (best fit line).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (I_m - I_c)^2}{N_t}} \quad (1)$$

$$MBE = \frac{\sum_{i=1}^{N} (I_m - I_c)}{N_t} \quad (2)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} (I_m - I_c)^2}{\sum_{i=1}^{N} (I_m - \overline{I_m})^2} \quad (3)$$

where I_m and I_c are the measured and estimated $PM_{2.5}$ concentrations, respectively, $\overline{I_m}$ is the average of the measured $PM_{2.5}$ and N the total number of measurements. $NSE = 1$ indicates an ideal model performance, while $NSE = 0$ shows model predictions as accurate as the mean of the observed data. Lower *RMSE* and *MBE* values correspond to better performance and to lower biases from the ML models.

4.3. Variable Importance

Variable importance statistic was used to analyse the contribution of each variable in $PM_{2.5}$ estimations. Since SVR is a kernel-based model, and we do not know the concrete form of its nonlinear mapping function, and the weight vector (ω) cannot be computed directly [128], it is complicated to analyse the variables importance statistic. On the other hand, there are two famous measures for RF, which are mean decrease accuracy (MDA) and mean decrease Gini (MDG). The MDG is based on Gini importance which measures the average gain by splits of a given variable, whilst MDA is based on out of bag (OOB) samples. In RF model, each tree is grown based on a bootstrap sample of the training data, and those data that were not used in the bootstrap sample are known as out of bag (OOB) samples [126]. The MDA measures the accuracy of the model losses by permuting each variable. This technique is considered as most efficient variable importance for random forest [129,131], and it was preferred in this study as less bias compared to Gini importance. The higher percentage value of the variable importance indicates higher influence of the corresponding variable to $PM_{2.5}$ estimations. In R script, we used the "varImp" function, which can automatically scale the importance scores in values between 0 and 100.

5. Results and Discussion

5.1. Descriptive Statistics

The descriptive statistics of the measured variables that are used in SVR and RF models for all stations in Malaysia are summarized in Table 1. The columnar AOD_{500} over the Malaysian sites during 2018–2019 exhibits a mean of 0.69, which is above the median value 0.46 due to many episodic aerosol events with AODs above 2, representing thick smoke plumes from extensive fires in Indonesia and Indochina [10,132,133]. The measured

$PM_{2.5}$ concentrations at the 65 examined sites follow a similar distribution with a higher mean (21.9 μg m^{-3}) than median (17.1 μg m^{-3}) and a maximum value of 230 μg m^{-3} (Table 1, Figure 3). These $PM_{2.5}$ levels are similar to those reported at several sites in Southeast Asia [86,134]. NO_2 and CO exhibit means of 5.2 ppb and 0.6 ppm, respectively, while tropospheric O_3 levels (25.2 ppb) are considered rather high with deleterious effects on human health [135,136].

Table 1. Statistical values for the measured parameters in all air-pollution monitoring sites.

	$PM_{2.5}$ (μg m^{-3})	AOD	SO_2 (ppb)	NO_2 (ppb)	O_3 (ppb)	CO (ppm)	WS (ms^{-1})	RH (%)	TEMP (°C)
Mean	21.86	0.69	1.2	5.23	25.2	0.60	1.73	66.58	30.70
Median	17.07	0.46	1.0	3.81	27.3	0.56	1.60	66.85	30.88
Stdev	19.15	0.68	0.9	6.1	15.1	0.28	1.02	10.35	2.31

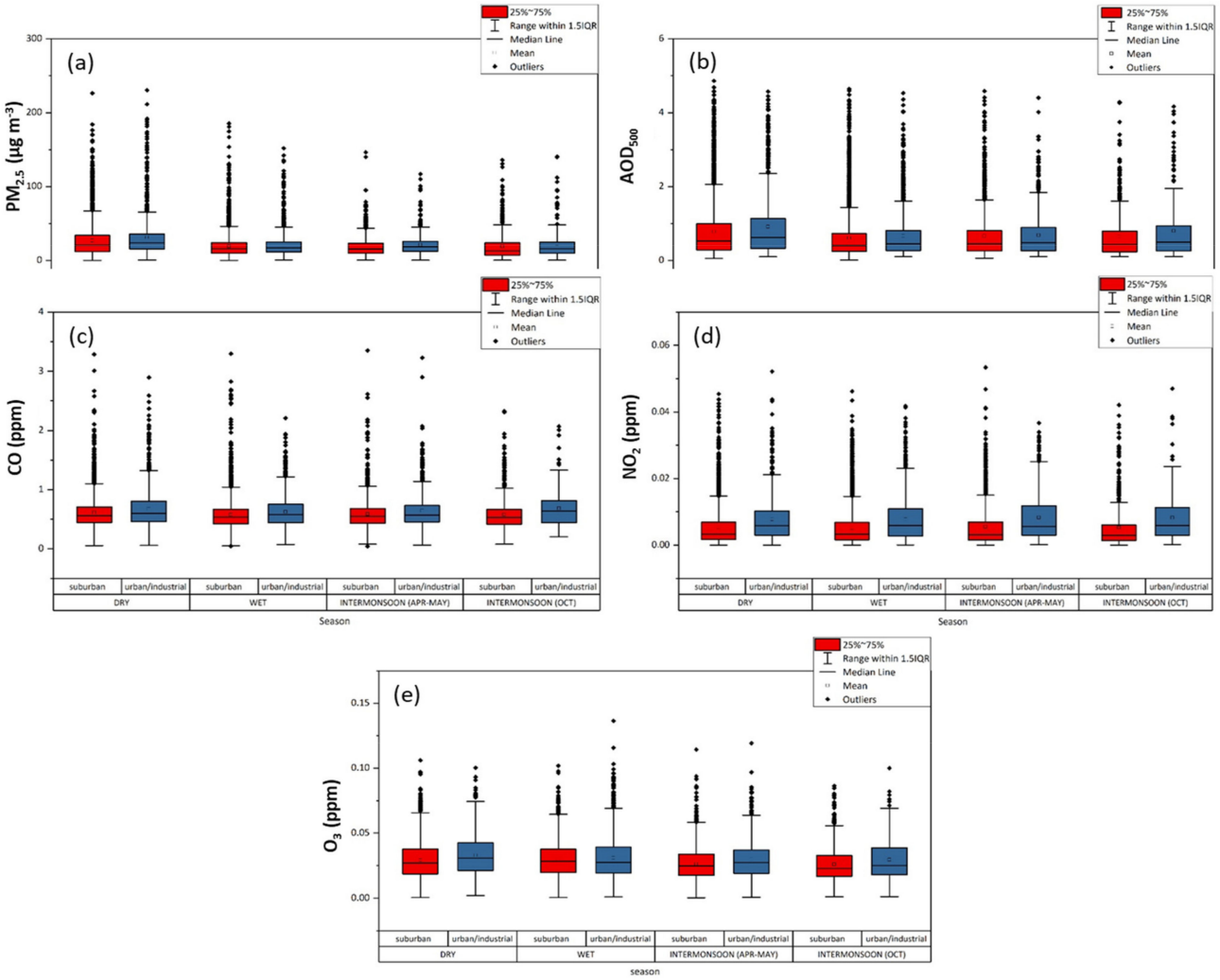

Figure 3. Box plots of measured $PM_{2.5}$ concentrations (**a**) Himawari-8 AOD_{500} (**b**), CO (**c**), NO_2 (**d**) and O_3 (**e**) at urban/industrial and suburban/rural sites in Malaysia in the dry, wet and inter-monsoon seasons.

The box-whisker plots for AOD_{500} and air pollutants, separated for urban/industrial and suburban/rural sites and for the four seasons, are shown in Figure 3. The analysis showed that the seasonal-mean $PM_{2.5}$ at the urban/industrial sites displayed higher values in all seasons. During the dry season, the maximum $PM_{2.5}$ levels were found to be 230.3 μg m^{-3} for the urban/industrial and 226.3 μg m^{-3} for the suburban sites, with

means of 31.26 μg m^{-3} and 26.38 μg m^{-3}, respectively. This is attributed to the prevailing southwest wind carrying biomass-burning aerosols from Indonesia due to extensive forest fires in this season [90]. However, the seasonal mean $PM_{2.5}$ concentrations do not notably differ in the other seasons (lying between 17.80 μg m^{-3} and 22.33 μg m^{-3} for both urban/industrial and suburban/rural sites), indicating a year-long $PM_{2.5}$ laden atmosphere across Malaysia. Regarding the Himawari-8 AOD, it is higher during the dry season, while slightly lower mean values (~0.6 to 0.8) were observed in the other seasons, with marginal differences between urban/industrial and suburban/rural sites (Figure 3b). The seasonal and site variations of the columnar AODs are in agreement with ground $PM_{2.5}$ concentrations, indicating that the industrial and traffic emissions are the main pollution sources for urban centres and surrounding areas. In general, heavy precipitation during the wet season only marginally reduced the aerosol levels since severe pollution episodes with $PM_{2.5}$ > 100 μg m^{-3} and AODs > 3 were also present. However, the columnar AOD displayed a very low correlation (R^2 = 0.09) with the surface $PM_{2.5}$, indicating (i) a significant aerosol loading aloft and (ii) different sources and temporal variability between surface $PM_{2.5}$ and AODs [137]. In addition, the mean concentrations of CO, NO_2 and O_3 (Figure 3c–e) are higher over the urban/industrial areas compared to suburban/rural sites in each season. Motor vehicle and power plants emissions are the major contributors to CO and NO_2 concentrations in Malaysia with about 95.7% and 66%, respectively [11], thus explaining the higher NO_2 and CO levels in urban/industrial areas. The stronger correlation was found between CO and $PM_{2.5}$ (R^2 = 0.33), revealing that the particulate pollution in Malaysia is mostly related to local sources of fossil fuel and biofuel combustion, which enhance CO emissions [86,138]. NO_2, which is mostly related to vehicular emissions, was negligibly associated with $PM_{2.5}$ concentrations (R^2 = 0.1), The overall mean SO_2 concentration was found to be 1.2 ppb, with slightly larger levels in the dry season, exhibiting means of 1.5 ppb and 1.3 ppb for the urban/industrial and suburban sites, respectively.

5.2. Models for $PM_{2.5}$ Estimation

In this study, seven models were developed for $PM_{2.5}$ estimations in Malaysia using ML techniques, namely, SVR and RF. These models were developed in order to better capture the remarkable spatial (between stations of different characteristics) and temporal (between seasons) variations of $PM_{2.5}$ and to examine the model's capability in representing the levels and evolution of $PM_{2.5}$. The developed models are:

Model 1: Overall model;
Model 2: Spatial model (urban/industrial);
Model 3: Spatial model (suburban/rural);
Model 4: Temporal model (dry season);
Model 5: Temporal model (wet season);
Model 6: Temporal model (inter-monsoon, April–May);
Model 7: Temporal model (inter-monsoon, October).

The models' inputs consist of AOD, SO_2, NO_2, CO, O_3, WS, WD, TEMP and RH, with the total number of the matching samples for all variables (overall model) to be 13,376 at the 65 monitoring stations from January 2018 to December 2019. The scatter plots between measured and predicted $PM_{2.5}$ concentrations via the SVR and RF models are shown in Figures 4 and 5 for the validation of the 7 developed models, while Supplementary Materials Table S2 summarizes the statistical indicators of R^2, *RMSE*, *MBE* and *NSE*, as well as the importance ranking, for each input variable at the 7 developed models. The statistical indicators correspond to the averaged values of each model for the given number of data (N).

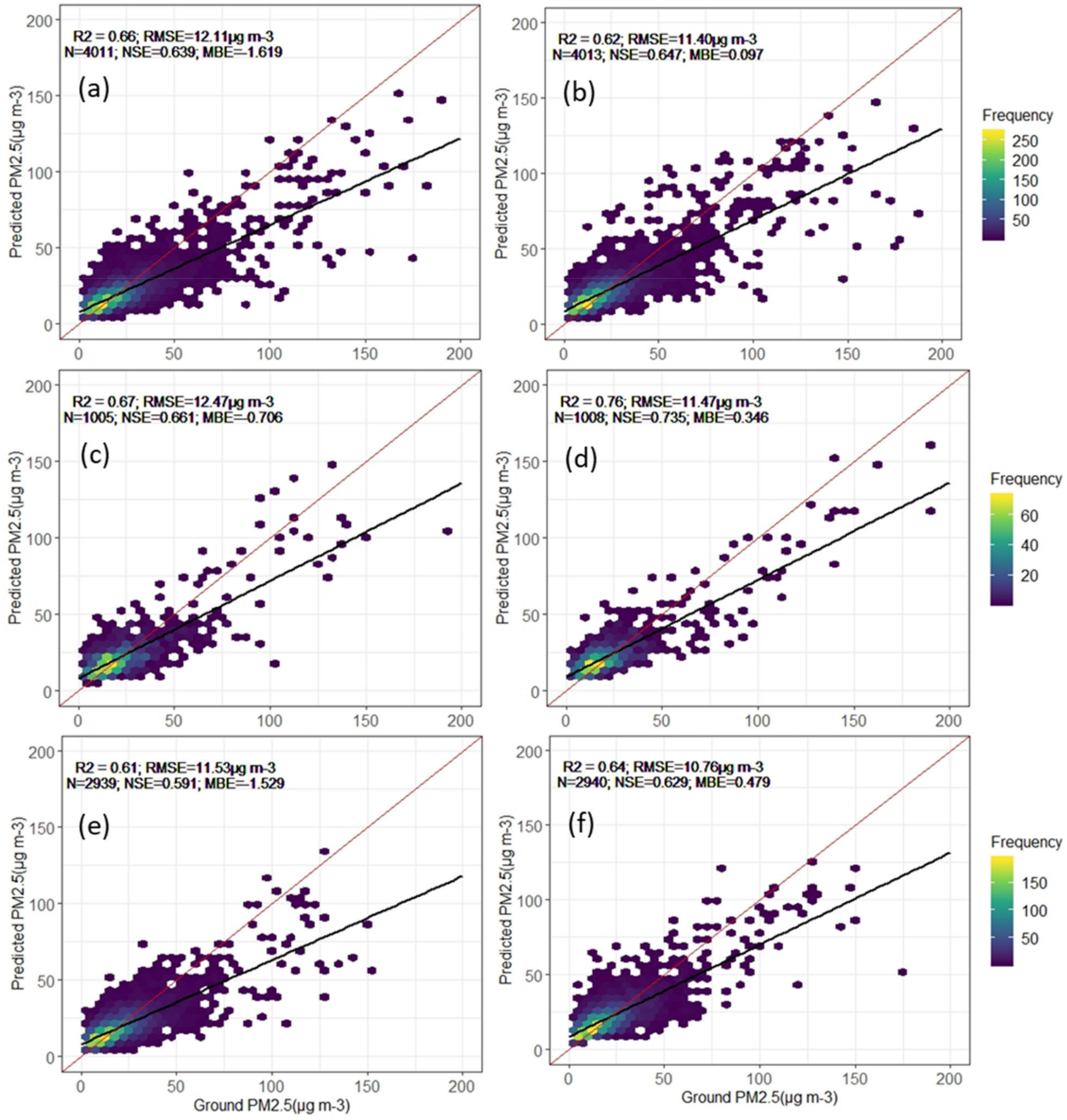

Figure 4. Validation of the predicted $PM_{2.5}$ concentrations against measured $PM_{2.5}$ using SVR (left) and RF (right) for overall Model 1 (**a**,**b**), urban/industrial Model 2 (**c**,**d**), suburban/rural Model 3 (**e**,**f**). Frequency indicates the density of data/count. The statistical indicators are presented as averaged values in each case.

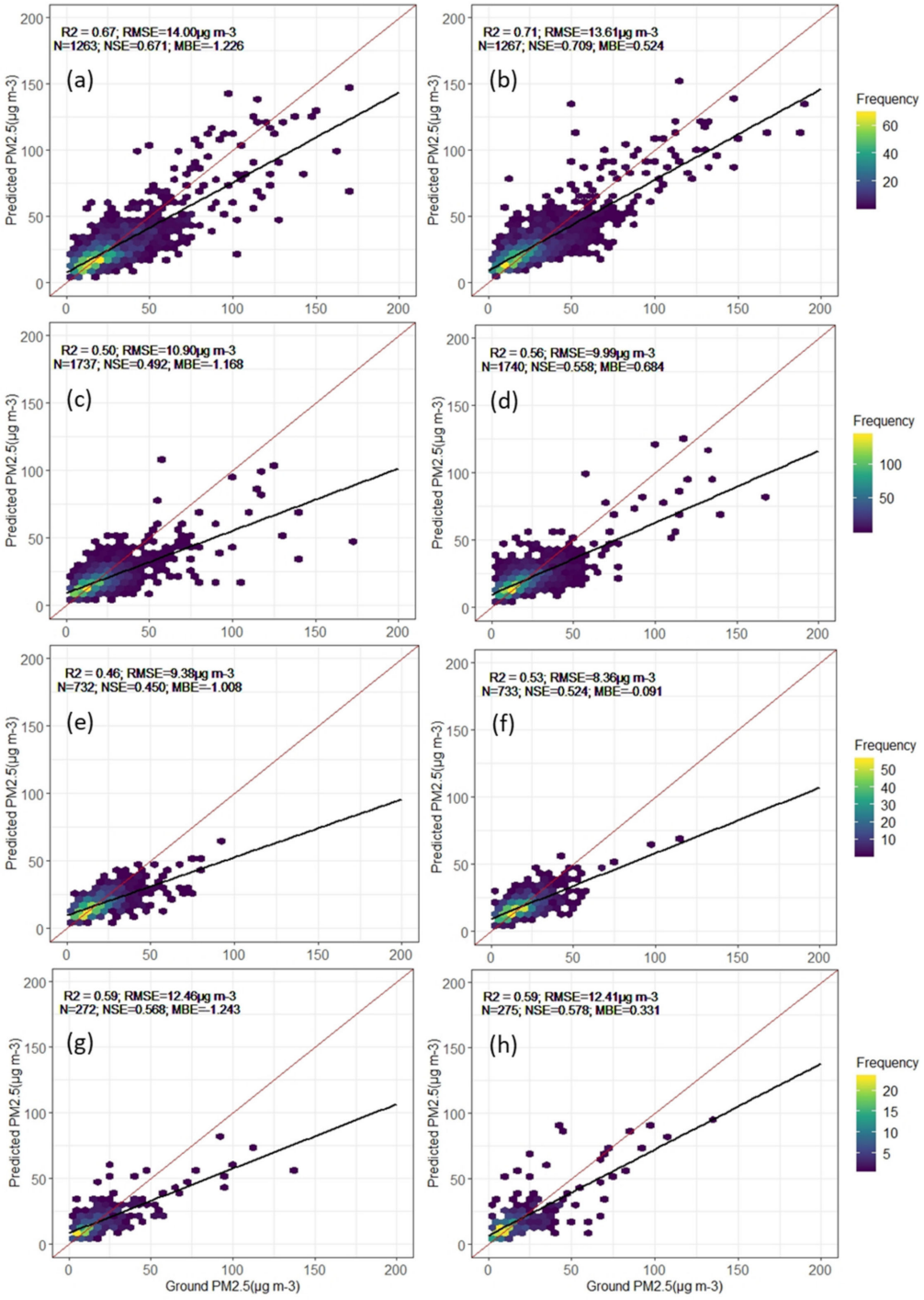

Figure 5. Validation of the predicted $PM_{2.5}$ concentrations against measured $PM_{2.5}$ using SVR (left) and RF (right) models in dry season (June–September) (**a**,**b**), wet season (November–March) (**c**,**d**), inter-monsoon (April–May) (**e**,**f**) and inter-monsoon (October) (**g**,**h**). Frequency indicates the density of data/count. The statistical indicators are presented as averaged values in each case.

The evaluation of the calibration datasets for the overall model 1 showed that the SVR model yielded comparable accuracy (R^2 = 0.69; *RMSE* = 10.62 µg m^{-3}, *NSE* = 0.679 and *MBE* = −1.392) to RF (R^2 = 0.66; *RMSE* = 11.28 µg m^{-3}, *NSE* = 0.939 and low *MBE* = 0.066) (Supplementary Materials Table S2). The validation dataset displayed also small differences between the two models and comparable statistics with the calibration datasets i.e., R^2 = 0.66 and *RMSE* = 12.11 µg m^{-3} for SVR and R^2 = 0.62, *RMSE* = 11.40 µg m^{-3} for RF (Figure 4).

Furthermore, other models for estimating $PM_{2.5}$ were also developed by splitting the entire datasets initially into two categories namely urban/industrial (3357 data points) and suburban/rural (N = 9798). The RF model performed slightly better compared to SVR for urban/industrial (model 2) and suburban (model 3) sites, with R^2 = 0.76, *RMSE* = 11.47 μg m^{-3}, *NSE* = 0.735 and R^2 = 0.67, *RMSE* = 12.47 μg m^{-3}, *NSE* = 0.661, respectively. Regarding the calibration datasets, the *NSE* values of all the RF models were usually above 0.9, indicating an accurate model's performance, compared to the SVR models, which exhibited NSE values in the range of 0.55 to 0.79 for the calibration datasets (Supplementary Materials Table S2). The model validation revealed that RF models performed slightly better than the SVR models for both locations. The validation of SVR and RF models for suburban sites was considered satisfactory with R^2 = 0.61, *RMSE* = 11.53 μg m^{-3} and R^2 = 0.64, *RMSE*= 10.76 μg m^{-3}, respectively (Figure 4b,c). Both models performed better at urban/industrial sites compared to the suburban sites, while all models underestimated the large $PM_{2.5}$ concentrations ($PM_{2.5}$ > 60 μg m^{-3}). For $PM_{2.5}$ below 50 μg m^{-3}, where the vast majority of the data points lie, the underestimated and overestimated data points are almost equal for all the models and the regression line tends to coincide with the 1-1 line (Figure 4).

In addition, four temporal models were also developed for each season in Malaysia, namely, dry season (June–September; Figure 5a,b), wet season (November–March; Figure 5c,d) and two inter-monsoon seasons (April–May; Figure 5e,f and October; Figure 5g,h). The total number of data for the temporal models is 4223, 5797, 2441 and 915 for the dry season, wet season, April-May and October, respectively. These datasets were randomly divided into calibration and validation groups, and the results of the statistical evaluations of the seasonal $PM_{2.5}$ predictions are included in Supplementary Materials Table S2 (models 4–7). The estimation accuracy of $PM_{2.5}$ concentrations via the SVR and RF models varied between the seasons, while it was relatively lower in the wet and inter-monsoon (April-May) seasons for both models. For instance, the R^2 values for the calibration dataset of the SVR model were 0.81, 0.62, 0.58 and 0.74 for the dry, wet seasons, April-May and October, respectively. The RF models exhibited slightly lower *RMSE* values compared to the SVR models, as 13.61 μg m^{-3}, 9.99 μg m^{-3}, 8.36 μg m^{-3} and 12.41 μg m^{-3} against 14.0, 10.9, 9.38 and 12.46 μg m^{-3} for dry, wet, April-May and October, respectively. Furthermore, RF models displayed higher *NSE* values compared to SVR, while the $PM_{2.5}$ underestimations from the SVR models (negative *MBE* values) are eliminated by using the RF models (Supplementary Materials Table S2; Figure 5).

The statistical evaluators of the developed models in Malaysia are mostly comparable to those found from multi-variate models including AOD and several meteorological parameters (Temp, RH, WS, Dew point, mixing height) for $PM_{2.5}$ estimations in Indian cities [139]. More recently, [39] developed a deep neural network consisted of recurrent layers for extracting the relationship between high-resolution (1 km) MODIS observations and PM_{10}, $PM_{2.5}$ concentrations in Tehran, Iran. The $PM_{2.5}$ and PM_{10} estimations resulted in *RMSE* values of 11.66 μg m^{-3} and 23.79 μg m^{-3}, respectively, comparable to the current results and previous *RMSE* values for PM_{10} estimations in Malaysia (11.61 μg m^{-3}; [18]) and Delhi (18.99 μg m^{-3}; [140]). MODIS-MAIAC AODs and columnar water vapor (CWV), along with meteorological parameters and land-use data, were included in a linear mixed effect model (LME) and a RF model for daily $PM_{2.5}$ estimations at high spatial resolution (1 km × 1 km) over the Indo-Gangetic Plains (IGP), India [118]. The RF model exhibited higher accuracy with R^2 = 0.87 and relative *RMSE* of 24.5%, compared to LME [118]. The spatial distributions of the R^2 (~0.6 to 0.9) and *RMSE* (~20 to 40 μg m^{-3}) values from $PM_{2.5}$ estimations across the IGP [118] were mostly comparable to those observed over Malaysia.

The hourly time series of the measured and predicted $PM_{2.5}$ concentrations via the SVR model, separately for the station characteristics and seasons and for the overall model, are shown in Figure 6. The results verify the good performance of the SVR model in predicting the $PM_{2.5}$ concentrations across Malaysia—RF performed slightly better with very similar results—also revealing an underestimation at the highest $PM_{2.5}$ values. However, the

model's underestimation in representing $PM_{2.5}$ peaks is not systematic, and in many cases, models reproduce satisfactorily the high $PM_{2.5}$ concentrations, even overestimating them (Figure 6).

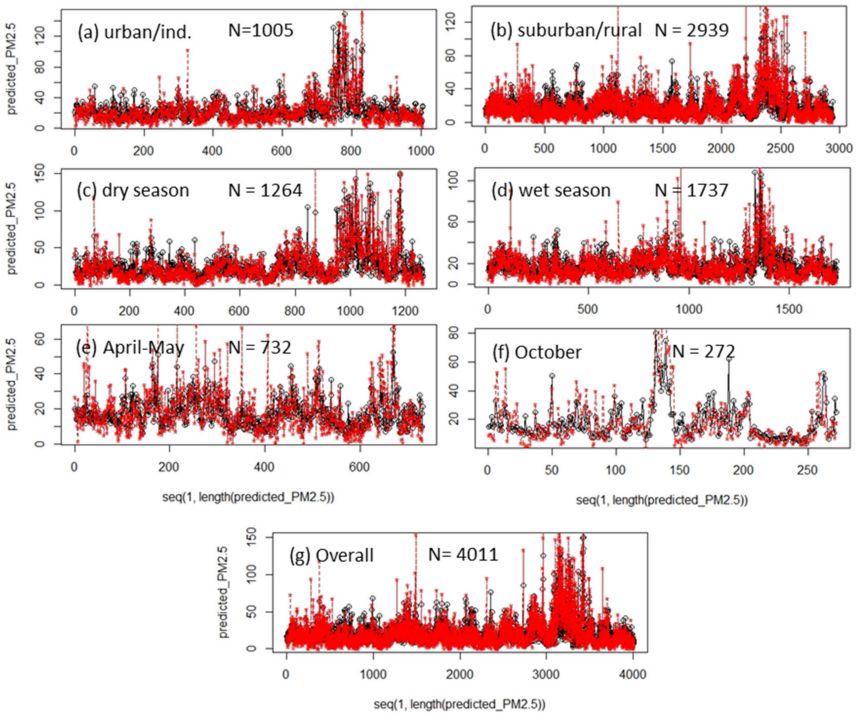

Figure 6. Time series (hourly) of measured (black) and predicted (red) $PM_{2.5}$ concentrations from the SVR model across Malaysia for urban/industrial (**a**), suburban/rural (**b**), dry season (**c**), wet season (**d**), April–May (**e**), October (**f**) and overall (**g**) models. The available data set for each group is mentioned in the panels. The X axis shows number of datasets.

The residual analysis for the model validation datasets revealed that the frequency of residuals approached the normal distribution peaking around zero for all the models (Supplementary Materials Figure S1). However, the frequency distributions were slightly shifted towards negative values, as the highest model underestimations may reach to −60%, but for very rare cases. In the vast majority of the cases, the predictions were quite accurate, indicating that the used ML models are satisfactory for $PM_{2.5}$ estimations in Malaysia.

Atmosphere is a complex system and composed by various substances like air molecules and solid and liquid particles of various sizes, shapes and chemical composition [141,142]. Therefore, combining many auxiliary data such as meteorological factors, aerosols, gases and land use allows for a better estimation of $PM_{2.5}$. To date, most PM prediction studies found that inclusion of meteorological factors has improved the PM estimations, because each meteorology parameter may modulate the PM concentrations in a different way [43,45,71]. Nowadays, ML techniques and RF models are frequently used in estimating PM concentrations at many regions around the world [39,116,119]. The statistical indicators from the models' calibration and validation in this study (Supplementary Materials Table S2; Figures 4 and 5) are mostly comparable to those presented in other studies using various methods and ML approaches for estimates of PM concentrations around the world (Table 2). It should be noted that this study obtained reasonable results at national scale without including land use information compared to previous works [128,143–145]. Besides that, the validation techniques may be also different, as for instance [143] used sample and site based 10 CV in order to assess the spatial performance, whilst our study only used sample based 10 CV since it can be used to reflect the overall predictive ability [130].

Table 2. Results from previous studies using machine learning techniques for PM estimations from space.

Author	Study Area	Input Data			Method	R/R^2	Accuracy
		Source of AOD	Other Parameters	Output			
[143]	China	MODIS AOD 10 km (Terra and Aqua)	RH, AT, WS, SP, PBLH, NDVI, population and road data	$PM_{2.5}$	Geoi-DBN	Sample based CV $R^2 = 0.88$ Site based CV $R^2 = 0.82$	Sample based CV *RMSE* = 13.03 μg m^{-3} Site based CV *RMSE* = 16.42 μg m^{-3}
[146]	China	MODIS AOD 3 km (Terra and Aqua)	Lat, long, month, RH, AT, WS, SP, PBLH	$PM_{2.5}$	GRNN	$R^2 = 0.89$	*RMSE* = 16.51 μg m^{-3}
[147]	China	MAIAC AOD 1 km	AT, AP, evaporation, precipitation, RH, sunshine duration and WS	$PM_{2.5}$	GW-GBM	Exclude missing AOD $R^2 = 0.74$ Include missing AOD $R^2 = 0.76$	Exclude missing AOD *RMSE* = 24.3 μg m^{-3} Include missing AOD *RMSE* = 23.0 μg m^{-3}
[128]	Cincinnati, OH, USA	MODIS AOD 3 km (Terra and Aqua)	Visibility, PBLH, TEMP, RH, total and rate precipitation, P, WS, WD, land cover, roadways, green space, spatiotemporal convolution layer	$PM_{2.5}$	RF	Overall $R^2 = 0.90$ Spatial $R^2 = 0.87$ Temporal $R^2 = 0.84$	Overall *RMSE* = 2.45 μg m^{-3} Spatial *RMSE* = 2.83 μg m^{-3} Temporal *RMSE* = 3.13 μg m^{-3}
[148]	British Columbia, Canada	MODIS AOD 3 km (Terra)	LST, humidity, vapour, NDVI, albedo from MODIS product. PBLH, WS. Elevation from SRTM	$PM_{2.5}$	MLR BRNN SVM LASSO MARS RF XGBoost Cubist	$R^2 = 0.22$ $R^2 = 0.31$ $R^2 = 0.30$ $R^2 = 0.24$ $R^2 = 0.31$ $R^2 = 0.49$ $R^2 = 0.46$ $R^2 = 0.48$	*RMSE* = 3.24 μg m^{-3} *RMSE* = 3.04 μg m^{-3} *RMSE* = 3.13 μg m^{-3} *RMSE* = 3.20 μg m^{-3} *RMSE* = 3.05 μg m^{-3} *RMSE* = 2.67 μg m^{-3} *RMSE* = 2.71 μg m^{-3} *RMSE* = 2.64 μg m^{-3}
[69]	BTH, China	MODIS AOD 10 km (Terra and Aqua)	PBLH, TEMP, SLP, humidity, WD and WS	$PM_{2.5}$	OR Rpart RF SVM	R = 0.73–0.76 R = 0.68–0.83 R = 0.69–0.84 R = 0.77–0.88	*RMSE* = 36.92–42.48 μg m^{-3} *RMSE* = 35.42–46.20 μg m^{-3} *RMSE* = 36.34–44.59 μg m^{-3} *RMSE* = 29.50–38.32 μg m^{-3}
[115]	East coast peninsular Malaysia	-	AT, RH, WS, GR, MSLP, rainfall, CO, NO_2, and SO_2	PM_{10}	MLR MLP RBF	$R^2 = 0.594$–0.706 $R^2 = 0.691$–0.794 $R^2 = 0.827$–0.929	*VIF* = 1.077–1.926 *RMSE* = 8.49–9.57 μg m^{-3} *RMSE* = 9.19–4.08 μg m^{-3}

Table 2. *Cont.*

Author	Study Area	Input Data			Method	R/R^2	Accuracy
		Source of AOD	Other Parameters	Output			
[141]	BTH, China	MODIS AOD 10 km (Aqua)	AT, RH, WS, WD and P	$PM_{2.5}$	MLR MARS SVR RSRF	R^2 = 0.733 R^2 = 0.776 R^2 = 0.850 R^2 = 0.843	*RMSE* = 33.016 μg m^{-3} *RMSE* = 30.180 μg m^{-3} *RMSE* = 24.745 μg m^{-3} *RMSE* = 25.320 μg m^{-3}
[149]	Wuhan, China	Himawari-8 AOD L3	MODIS NDVI, RH, AT, WS, SP, PBLH, DEM	$PM_{2.5}$	DL	R^2 = 0.850	*RMSE* = 9.303 μg m^{-3}
[144]	China	MAIAC AOD 1 km	TEMP, total precipitation, evaporation, PBLH, RH, SP, WS, WD, MODIS Land use Cover, NDVI, DEM	$PM_{2.5}$	RF STRF	R^2 = 0.98 R^2 = 0.98	*RMSE* = 6.40 μg m^{-3} *RMSE* = 5.57 μg m^{-3}
[150]	Shenzhen, China	MAIAC AOD 1 km	EWS and RH	$PM_{2.5}$	RF IRF	R^2 = 0.88 R^2 = 0.91	*RMSE* = 4.3 μg m^{-3} *RMSE* = 3.66 μg m^{-3}
[145]	East Asia (Eastern China, Korean Peninsular and Japan)	GOCI, GEOS-Chem	NDVI, urban ratio, DEM, precipitation, AT, ST, dew point temperature, RH, max WS, visibility, PBLH, SP, solar radiation, road density, population density	PM_{10} $PM_{2.5}$	RF	R^2 = 0.88 R^2 = 0.90	*RMSE* = 26.9 μg m^{-3} *RMSE* = 15.77 μg m^{-3}
[70]	Greater London	MAIAC AOD 1 km	Population density, cloudiness, barometric pressure, WD, WS, dew point temperature, land use variable (type, distance to water, airport, PBLH, NDVI, traffic count, elevation etc)	$PM_{2.5}$	GBM RF Deep NN KNN ensemble model	Overall model R^2 = 0.826 R^2 = 0.830 R^2 = 0.793 R^2 = 0.791 R^2 = 0.828	Overall model *RMSE* = 4.331 μg m^{-3} *RMSE* = 4.278 μg m^{-3} *RMSE* = 4.728 μg m^{-3} *RMSE* = 4.721 μg m^{-3} *RMSE* = 4.231 μg m^{-3}
[151]	Guwahati, India	–	CO, NO_2, SO_2, AT, RH, WS, rainfall	PM_{10}	MLR MLP CART	R^2 = 0.61–0.68 R^2 = 0.64–0.69 R^2 = 0.52–0.63	*RMSE* =29.31–31.99 μg m^{-3} *RMSE* =31.02–31.74 μg m^{-3} *RMSE* = 39.98–41.24 μg m^{-3}

List of abbreviations: 1. Study area: BTH (Beijing-Tianjin-Hebei). 2. Sensor: GOCI (Geostationary Ocean Color Imager), MAIAC (Multiangle Implementation of Atmospheric Correction), MODIS (Moderate Resolution Imaging Spectroradiometer), SRTM (*Shuttle Radar Topography Mission),* 3. Parameter: AT (air temperature), CO (carbon monoxide), EWS (extreme wind speed), GR (global radiation), Lat (Latitude), Long (longitude), LST (land surface temperature), MSLP (mean sea level pressure), NDVI (Normalized difference vegetation index), NO_2 (nitrogen dioxide), P (pressure), PBLH (planetary boundary layer height), RH (relative humidity), SF (surface temperature), SLP (sea level pressure), SO_2 (sulphur dioxide), SP (surface pressure), ST (surface temperature), TEMP (temperature), WD (wind direction), WS (wind speed). 4. Method: BRNN (Bayesian Regularized Neural Networks), CART (Classification and Regression Trees), Cubist (rule based tree model), DL (Deep Learning), DNN (deep neural network), GBM (Gradient Boosting Machine), Geoi-DBN (Geo-intelligent Deep Belief Network), GRNN (generalized regression neural network), GW-GBM (Geographically- Weighted Gradient Boosting Machine), IRF (improved random forest), KNN (k-nearest neighbour), LASSO (Least Absolute Shrinkage and Selection Operator), MARS (Multivariate Adaptive Regression Splines), MLP (multilayer perceptron), MLR (Multiple Linear Regression), OR (orthogonal regression), RBF (radial basis function), RF (random forest), Rpart (regression tree), RSRF (hybrid remote sensing and random forest), STRF (space-time random forest), SVM (Support Vector Machines), SVR (support vector regression), XGBoost (eXtreme Gradient Boosting). 5. Accuracy: *RMSE* (root means square error), VIF (variance inflation factor).

Determining the strength of the correlation between $PM_{2.5}$ and all the parameters used for its prediction is very important because it can indirectly portray the pollution process and the source of the pollution. The results of the variable importance analysis for the RF model have been included in Supplementary Materials Table S2 and are shown in Figure 7. For the overall model 1, CO is the highest contributor to the $PM_{2.5}$ estimations, similar to the other models, as discussed above, and is followed by AOD, O_3, NO_2, SO_2 and the meteorological parameters. This is because both $PM_{2.5}$ and CO are originated from common sources in Malaysia like biomass burning and traffic which enhance the CO emissions [86,138]. Besides that, CO may stay in the atmosphere for a long period (weeks or months), being able to get transported in high concentrations from biomass-burning areas in Indonesia [152]. Parameters with the least importance in $PM_{2.5}$ estimations are RH, WD, TEMP and finally the WS with a zero score (Figure 7a). Similar to the overall model (model 1), the contributions of the meteorological parameters were relatively weak in the spatial models as well, which also revealed CO, AOD and O_3 as the most important variables (Figure 7b,c). CO remains the most important predictor in the seasonal models as well, with minimal contributions from the meteorological variables (Figure 7d–g). The WS and WD have minimal contributions, in agreement with [81], who found that both parameters were not significantly correlated with $PM_{2.5}$ in the Klang Valley region in Peninsular Malaysia. Although AOD usually exhibited a high correlation with $PM_{2.5}$ [153], in our case, there was not a direct association with $PM_{2.5}$ concentrations, implying complicated pollution conditions in the vertical layer over Malaysia [141]. Generally, $PM_{2.5}$ is affected more by gaseous pollutants and not so much by the columnar AOD (missing values due to cloud cover and elevated aerosol layers) and meteorological conditions since Malaysia has rather uniform weather conditions throughout the year. Therefore, influence of the meteorological parameters is minimal towards seasonally changing $PM_{2.5}$ concentrations. However, previous studies in Malaysia showed that the meteorological parameters affected the coarse particles, e.g., PM_{10} concentrations [81], indicating a meteorological-dependent character of the coarse-mode aerosols. In our previous study [18], we found that the estimations of PM_{10} concentrations based on satellite AODs were significantly improved after inclusion of the meteorological parameters in the model.

Similar to the current results, [154] also found that the parameters with the highest importance in predicting $PM_{2.5}$ concentrations are CO, NO_2, SO_2 and AOD. Inclusion of the pollutant gases improved the performance of their RF model from $R^2 = 0.69$, $RMSE = 41.63$ μg m^{-3} to $R^2 = 0.81$, $RMSE = 32.74$ μg m^{-3}, in a similar way as in the present study. These results were also in agreement with [65] who found that CO was the most important variable that explained 20.65% of the variation in estimated $PM_{2.5}$ concentrations in Xi'an, China, exhibiting a strong correlation with AOD. Furthermore, [81] studied the $PM_{2.5}$ composition in Klang Valley, Malaysia, and concluded that CO, NO_2, NO and SO_2 mostly affected the $PM_{2.5}$ concentrations.

This is the first study conducted in Malaysia aiming to estimate the $PM_{2.5}$ concentrations based on machine-learning techniques. The satisfactory accuracy of the estimates, despite the biases and challenges in representing $PM_{2.5}$ pollution episodes, is especially important for the development of models aiming to systematically monitor $PM_{2.5}$ over the country, especially at remote areas with unavailability of measurements. However, the only 65 operational stations are still insufficient to cover the whole Malaysian territory with an area of 330,290 km^2. Establishing more air quality monitoring stations is very costly, and a certain station is only capable to satisfactorily represent the pollutant concentrations within a radius of about 15 km [155]. Alternatively, remote sensing data encourages more studies to be conducted on atmospheric particulates and air quality, since satellite technology considers AOD as a key predictor of PM over a large area [156]. This would also help in evaluating the influence of the local/regional emissions from anthropogenic activities against those attributed to natural causes or long-range transported aerosols, mostly smoke, from Indonesia and other parts in northern Indochina.

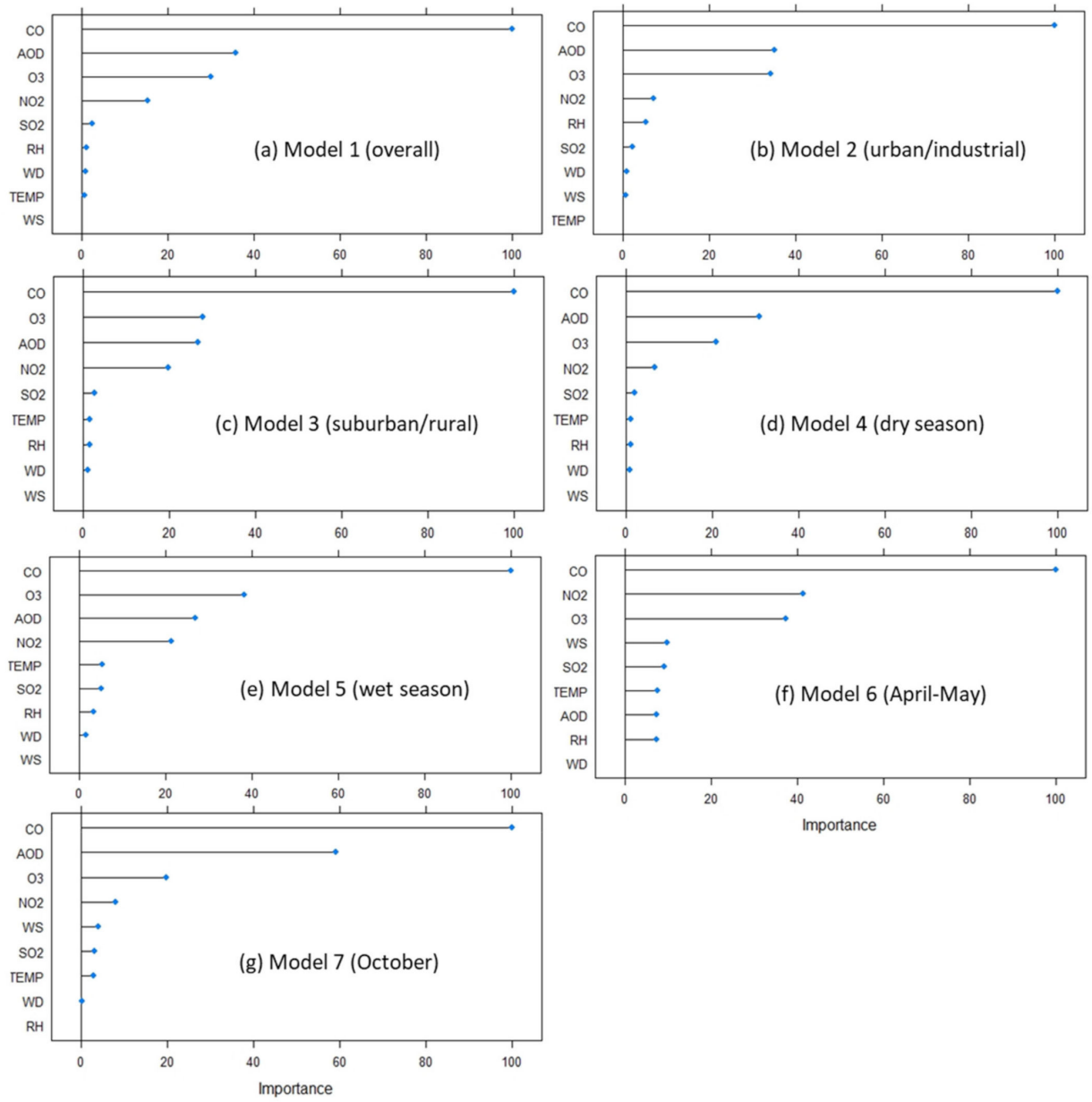

Figure 7. Variable importance analysis for the RF Models 1–7 (**a**–**g**). Y-axis indicates the predictor for $PM_{2.5}$ estimations, and x-axis indicates the importance scores between 0 and 100.

6. Conclusions

The current study developed new machine-learning models, namely, Random Forest (RF) and Support Vector Regression (SVR), to estimate $PM_{2.5}$ concentrations across Malaysia for the first time covering the years 2018 and 2019. Satellite (Himawari-8) AOD, ground-measured air pollutants (NO_2, CO, SO_2, O_3) and meteorological parameters (temperature, relative humidity, wind speed and direction) were used as input variables. Due to the high spatial (between stations with different characteristics like urban/industrial and suburban/rural) and temporal (between seasons) evolution of the $PM_{2.5}$ levels across Malaysia, seven sub-models were developed separately for the different sites (urban/industrial, suburban/rural) and seasons (dry, wet and two inter-monsoons (April-May and October)), and one overall model. Of the available dataset, 70% was randomly selected for the model calibration, and the remaining 30% for the model validation. The $PM_{2.5}$ predictions of each model are compared to those measured at 65 air pollution monitoring stations, using standard statistical estimators.

For the overall model, SVR calibration performed slightly better than RF with $R^2 = 0.69$ and *RMSE* = 10.62 μg m^{-3} against measured $PM_{2.5}$ concentrations. Whilst for the spatial models, the RF validation performed slightly better than SVR, with statistical indicators of $R^2 = 0.76$, *RMSE* = 11.47 μg m^{-3} for urban/industrial, and $R^2 = 0.64$, *RMSE* = 10.76 μg m^{-3} for the suburban/rural sites. Therefore, both RF and SVR models displayed slightly higher performance for $PM_{2.5}$ estimations at urban/industrial sites with higher levels of AOD and air pollutants. Furthermore, the estimation accuracy of SVR and RF models was lower in the wet (November–March) and inter-monsoon (April–May) seasons compared to the dry (June–September) season. Based on the model accuracy and variable importance analysis, CO was always the most influential predictor variable for $PM_{2.5}$ estimations in Malaysia, followed by AOD, O_3, NO_2, SO_2 and meteorological parameters but with different order depending on the dataset and model. An important finding was the very weak correlation and contribution of the meteorological variables to $PM_{2.5}$ estimations. Furthermore, very low correlation was found between $PM_{2.5}$ and columnar AOD, indicating that surface pollution followed a different temporal pattern than AOD and the presence of a significant aerosol layer aloft due to transported smoke plumes from wildfires in southeast Asia. The current results showed that the use of machine-learning techniques for $PM_{2.5}$ estimations over Malaysia was promising as these models can satisfactorily represent the values and temporal evolution of $PM_{2.5}$ concentrations over both urban/industrial and suburban/rural sites although the underestimation of the highest $PM_{2.5}$ levels. In a next step, gaseous pollutants from satellite remote sensing observations will be included in ML approaches in order to estimate $PM_{2.5}$ concentrations over large areas, aiming to cover the whole Malaysian territory.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/app11167326/s1, Figure S1: Residual analysis (residuals = predicted $PM_{2.5}$—measured $PM_{2.5}$) from SVR and RF for the developed Models 1–7. The fitted curve represents the normal distribution, Table S1: T Parameters/functions used for the SVR model, Table S2: Coefficient of determination (R2), *RMSE*, *MBE* and NSE values using SVR and RF models for $PM_{2.5}$ estimations in Malaysia. The statistical indicators are presented as averaged values for each model and Number of samples.

Author Contributions: Investigation, N.A.F.K.Z., K.D.K.; methodology, N.A.F.K.Z., K.D.K., D.G.K.; data curation, N.A.F.K.Z., K.D.K., D.G.K., M.T.L.; formal analysis, N.A.F.K.Z., K.D.K., D.G.K.; writing—original draft, N.A.F.K.Z., K.D.K.; conceptualization, K.D.K.; writing—review and editing, N.A.F.K.Z., K.D.K., D.G.K., M.T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Ministry of Education, Malaysia, via the Fundamental Research Grant (FRGS/1/2019/WAB05/UTM/02/3) and WNI WXBUNKA Foundation, Japan, via research grant R.J130000.7352.4B406.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Himawari-8 data can be accessed via http://www.eorc.jaxa.jp/ptree/index.html. Ground-based pollution data can be accessed after request.

Acknowledgments: The authors would like to thank the Japan Aerospace Exploration Agency (JAXA) and Department of Environment, Malaysia, for providing the Himawari-8 AOD data and surface air pollutant data, respectively. D.G.K. acknowledges support of the project PANACEA (PANhellenic infrastructure for Atmospheric Composition and climatE change; MIS 5021516), under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). Valuable comments from two anonymous reviewers are highly appreciated.

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

References

1. Ranjan, A.K.; Patra, A.K.; Gorai, A. A Review on Estimation of Particulate Matter from Satellite-Based Aerosol Optical Depth: Data, Methods, and Challenges. *Asia Pac. J. Atmos. Sci.* **2020**, *57*, 679–699. [CrossRef]
2. Jain, S.; Sharma, S.K.; Vijayan, N.; Mandal, T.K. Seasonal characteristics of aerosols ($PM_{2.5}$ and PM_{10}) and their source apportionment using PMF: A four year study over Delhi. India. *Environ. Pollut.* **2020**, *262*, 114337. [CrossRef]
3. Pani, S.K.; Wang, S.H.; Lin, N.H.; Chantara, S.; Lee, C.T.; Thepnuan, D. Black carbon over an urban atmosphere in northern peninsular Southeast Asia: Characteristics, source apportionment, and associated health risks. *Environ. Pollut.* **2020**, *259*, 113871. [CrossRef]
4. WHO. WHO Global Ambient Air Quality Database. 2020. Available online: https://www.who.int/data/gho/data (accessed on 21 October 2020).
5. Latif, M.T.; Othman, M.; Idris, N.; Juneng, L.; Abdullah, A.M.; Hamzah, W.P.; Khan, M.F.; Nik Sulaiman, N.M.; Jewaratnam, J.; Aghamohammadi, N.; et al. Impact of regional haze towards air quality in Malaysia: A review. *Atmos. Environ.* **2018**, *177*, 28–44. [CrossRef]
6. Jamalani, M.; Abdullah, A.; Azid, A.; Ramli, M.; Baharudin, M.; Chng, L.; Elhadi, R.; Yusof, K.K.; Gnadimzadeh, A. PM 10 emission inventory of industrial and road transport vehicles in Klang Valley, Peninsular Malaysia. *J. Fundam. Appl. Sci.* **2018**, *10*, 313–324.
7. Chang, S.-C. Atmospheric impacts of Indonesian fire emissions: Assessing remote sensing data and air quality during 2013 Malaysian haze. *Procedia Environ. Sci.* **2016**, *36*, 176–179.
8. Gautam, R.; Hsu, N.C.; Eck, T.F.; Holben, B.N.; Janjai, S.; Jantarach, T.; Tsay, S.-C.; Lau, W.K. Characterization of aerosols over the Indochina peninsula from satellite-surface observations during biomass burning pre-monsoon season. *Atmos. Environ.* **2013**, *78*, 51–59. [CrossRef]
9. Wang, S.-H.; Welton, E.J.; Holben, B.N.; Tsay, S.-C.; Lin, N.-H.; Giles, D.; Stewart, S.A.; Janjai, S.; Nguyen, X.A.; Hsiao, T.-C. Vertical distribution and columnar optical properties of springtime biomass-burning aerosols over Northern Indochina during 2014 7-SEAS campaign. *Aerosol Air Qual. Res.* **2015**, *15*, 2037–2050. [CrossRef]
10. Reid, J.S.; Hyer, E.J.; Johnson, R.S.; Holben, B.N.; Yokelson, R.J.; Zhang, J.; Campbell, J.R.; Christopher, S.A.; Di Girolamo, L.; Giglio, L. Observing and understanding the Southeast Asian aerosol system by remote sensing: An initial review and analysis for the Seven Southeast Asian Studies (7SEAS) program. *Atmos. Res.* **2013**, *122*, 403–468. [CrossRef]
11. Official Portal of Department of Environment. *Environmental Quality Report 2019*; DOE: Putrajaya, Malaysia, 2019.
12. Hyer, E.J.; Reid, J.S.; Prins, E.M.; Hoffman, J.P.; Schmidt, C.C.; Miettinen, J.I.; Giglio, L. Patterns of fire activity over Indonesia and Malaysia from polar and geostationary satellite observations. *Atmos. Res.* **2013**, *122*, 504–519. [CrossRef]
13. Khan, F.; Latif, M.T.; Juneng, L.; Amil, N.; Mohd Nadzir, M.S.; Syedul Hoque, H.M. Physicochemical factors and sources of particulate matter at residential urban environment in Kuala Lumpur. *J. Air Waste Manag. Assoc.* **2015**, *65*, 958–969. [CrossRef]
14. Kanniah, K.D.; Zaman, N.A.F.K. Remotely sensed particulate matter estimation in malaysia during the biomass burning season in southeast Asia. In *Biomass Burning in South and Southeast Asia Impacts on the Biosphere*; Vadrevu, K.P., Ohara, T., Justice, C., Eds.; CRC Press Taylor and Francis: London, UK, 2021; Volume 2.
15. Department of Statistics Malaysia, Official Portal. Statistics on Causes of Death, Malaysia, 2019. Available online: https://www.dosm.gov.my (accessed on 16 December 2020).
16. Dominici, F.; Peng, R.D.; Bell, M.L.; Pham, L.; McDermott, A.; Zeger, S.L.; Samet, J.M. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **2006**, *295*, 1127–1134. [CrossRef] [PubMed]
17. Lary, D.; Lary, T.; Sattler, B. Using machine learning to estimate global $PM_{2.5}$ for environmental health studies. *Environ. Health Insights* **2015**, *9*, EHI-S15664. [CrossRef] [PubMed]
18. Zaman, N.A.; Kanniah, K.D.; Kaskaoutis, D.G. Estimating Particulate Matter using satellite based aerosol optical depth and meteorological variables in Malaysia. *Atmos. Res.* **2017**, *193*, 142–162. [CrossRef]
19. Othman, M.; Latif, M.T.; Jamhari, A.A.; Abd Hamid, H.H.; Uning, R.; Khan, M.F.; Nadzir, M.S.M.; Sahani, M.; Wahab, M.I.A.; Chan, K.M. Spatial distribution of fine and coarse particulate matter during a southwest monsoon in Peninsular Malaysia. *Chemosphere* **2021**, *262*, 127767. [CrossRef]
20. Stowe, L.L.; Ignatov, A.M.; Singh, R.R. Development, validation, and potential enhancements to the second-generation operational aerosol product at the National Environmental Satellite, Data, and Information Service of the National Oceanic and Atmospheric Administration. *J. Geophys. Res. Atmos.* **1997**, *102*, 16923–16934. [CrossRef]
21. Ignatov, A.; Sapper, J.; Cox, S.; Laszlo, I.; Nalli, N.R.; Kidwell, K.B. Operational Aerosol Observations (AEROBS) from AVHRR/3 On Board NOAA-KLM Satellites. *J. Atmos. Ocean. Technol.* **2004**, *21*, 3–26. [CrossRef]
22. Kahn, R.; Gaitley, B.; Martonchik, J.; Diner, D.; Crean, K.; Holben, B. MISR global aerosol optical depth validation based on two years of coincident data AERONET observations. *J. Geophys. Res.* **2004**, *109*. [CrossRef]
23. Vidot, J.; Santer, R.; Aznay, O. Evaluation of the MERIS aerosol product over land with AERONET. *Atmos. Chem. Phys.* **2008**, *8*, 7603–7617. [CrossRef]
24. Schmid, J. The SEVIRI Instrument. In Proceedings of the 2000 EUMETSAT Meteorological Satellite Data User's Conference, Bologna, Italy, 29 May–2 June 2000.
25. Remer, L.A.; Kaufman, Y.; Tanré, D.; Mattoo, S.; Chu, D.; Martins, J.V.; Li, R.-R.; Ichoku, C.; Levy, R.; Kleidman, R. The MODIS aerosol algorithm, products, and validation. *J. Atmos. Sci.* **2005**, *62*, 947–973. [CrossRef]

26. Levy, R.; Remer, L.; Kleidman, R.; Mattoo, S.; Ichoku, C.; Kahn, R.; Eck, T. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. *Atmos. Chem. Phys.* **2010**, *10*, 10399–10420. [CrossRef]
27. Munchak, L.; Levy, R.; Mattoo, S.; Remer, L.; Holben, B.; Schafer, J.; Hostetler, C.; Ferrare, R. MODIS 3 km aerosol product: Applications over land in an urban/suburban region. *Atmos. Meas. Tech.* **2013**, *6*, 1747–1759. [CrossRef]
28. Sever, L.; Alpert, P.; Lyapustin, A.; Wang, Y.; Chudnovsky, A. An example of aerosol pattern variability over bright surface using high resolution MODIS MAIAC: The eastern and western areas of the Dead Sea and environs. *Atmos. Environ.* **2017**, *165*, 359–369. [CrossRef]
29. Remer, L.; Mattoo, S.; Levy, R.; Munchak, L. MODIS 3km aerosol product: Algorithm and global perspective. *Atmos. Meas. Tech.* **2013**, *6*, 1829–1844. [CrossRef]
30. Kanniah, K.D.; Kaskaoutis, D.G.; San Lim, H.; Latif, M.T.; Kamarul Zaman, N.A.F.; Liew, J. Overview of atmospheric aerosol studies in Malaysia: Known and unknown. *Atmos. Res.* **2016**, *182*, 302–318. [CrossRef]
31. Zaman, N.A.; Kanniah, K.D.; Kaskaoutis, D.G. Satellite Data for Upscalling Urban Air Pollution in Malaysia. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Kuala Lumpur, Malaysia, 24–25 April 2018; p. 012036.
32. Kanniah, K.D.; Lim, H.Q.; Kaskaoutis, D.G.; Cracknell, A.P. Investigating aerosol properties in Peninsular Malaysia via the synergy of satellite remote sensing and ground-based measurements. *Atmos. Res.* **2014**, *138*, 223–239. [CrossRef]
33. Chudnovsky, A.A.; Lee, H.J.; Kostinski, A.; Kotlov, T.; Koutrakis, P. Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). *J. Air Waste Manag. Assoc.* **2012**, *62*, 1022–1031. [CrossRef]
34. Xiao, Q.; Zhang, H.; Choi, M.; Li, S.; Kondragunta, S.; Kim, J.; Holben, B.; Levy, R.; Liu, Y. Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia. *Atmos. Chem. Phys.* **2016**, *16*. [CrossRef]
35. Yang, J.; Zhang, Z.; Wei, C.; Lu, F.; Guo, Q. Introducing the new generation of Chinese geostationary weather satellites, Fengyun-4. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 1637–1658. [CrossRef]
36. Bessho, K.; Date, K.; Hayashi, M.; Ikeda, A.; Imai, T.; Inoue, H.; Kumagai, Y.; Miyakawa, T.; Murata, H.; Ohno, T. An introduction to Himawari-8/9—Japan's new-generation geostationary meteorological satellites. *J. Meteorol. Soc. Jpn. Ser. II* **2016**, *94*, 151–183. [CrossRef]
37. Zhang, W.; Xu, H.; Zhang, L. Assessment of Himawari-8 AHI Aerosol Optical Depth Over Land. *Remote Sens.* **2019**, *11*, 1108. [CrossRef]
38. Yang, X.; Zhao, C.; Luo, N.; Zhao, W.; Shi, W.; Yan, X. Evaluation and Comparison of Himawari-8 L2 V1.0, V2.1 and MODIS C6.1 aerosol products over Asia and the oceania regions. *Atmos. Environ.* **2020**, *220*, 117068. [CrossRef]
39. Imani, M. Particulate matter ($PM_{2.5}$ and PM_{10}) generation map using MODIS Level-1 satellite images and deep neural network. *J. Environ. Manag.* **2021**, *281*, 111888. [CrossRef]
40. Xu, X.; Zhang, C.; Liang, Y. Review of Satellite-driven Statistical Models $PM_{2.5}$ Concentration Estimation with Comprehensive Information. *Atmos. Environ.* **2021**, *256*, 118302. [CrossRef]
41. 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*, 2003. [CrossRef]
42. Engel-Cox, J.A.; Holloman, C.H.; Coutant, B.W.; Hoff, R.M. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmos. Environ.* **2004**, *38*, 2495–2509. [CrossRef]
43. Benas, N.; Beloconi, A.; Chrysoulakis, N. Estimation of urban PM_{10} concentration, based on MODIS and MERIS/AATSR synergistic observations. *Atmos. Environ.* **2013**, *79*, 448–454. [CrossRef]
44. Chitranshi, S.; Sharma, S.P.; Dey, S. Satellite-based estimates of outdoor particulate pollution (PM_{10}) for Agra City in northern India. *Air Qual. Atmos. Health* **2015**, *8*, 55–65. [CrossRef]
45. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res. Atmos.* **2009**, *114*. [CrossRef]
46. Schaap, M.; Apituley, A.; Timmermans, R.; Koelemeijer, R.; Leeuw, G.D. Exploring the relation between aerosol optical depth and PM 2.5 at Cabauw, the Netherlands. *Atmos. Chem. Phys.* **2009**, *9*, 909–925. [CrossRef]
47. Liu, Y.; Park, R.J.; Jacob, D.J.; Li, Q.; Kilaru, V.; Sarnat, J.A. Mapping annual mean ground-level $PM_{2.5}$ concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States. *J. Geophys. Res. Atmos.* **2004**, *109*. [CrossRef]
48. 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*. [CrossRef]
49. Engel-Cox, J.; Oanh, N.T.; van Donkelaar, A.; Martin, R.V.; Zell, E. Toward the next generation of air quality monitoring: Particulate Matter. *Atmos. Environ.* **2013**, *80*, 584–590. [CrossRef]
50. Crouse, D.L.; Philip, S.; Van Donkelaar, A.; Martin, R.V.; Jessiman, B.; Peters, P.A.; Weichenthal, S.; Brook, J.R.; Hubbell, B.; Burnett, R.T. A new method to jointly estimate the mortality risk of long-term exposure to fine particulate matter and its components. *Sci. Rep.* **2016**, *6*, 18916. [CrossRef] [PubMed]
51. Lee, H.; Liu, Y.; Coull, B.; Schwartz, J.; Koutrakis, P. A novel calibration approach of MODIS AOD data to predict PM 2.5 concentrations. *Atmos. Chem. Phys. Discuss.* **2011**, *11*, 7991–8002. [CrossRef]
52. Kloog, I.; Koutrakis, P.; Coull, B.A.; Lee, H.J.; Schwartz, J. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **2011**, *45*, 6267–6275. [CrossRef]

53. Xie, Y.; Wang, Y.; Zhang, K.; Dong, W.; Lv, B.; Bai, Y. Daily estimation of ground-level $PM_{2.5}$ concentrations over Beijing using 3 km resolution MODIS AOD. *Environ. Sci. Technol.* **2015**, *49*, 12280–12288. [CrossRef]
54. Beloconi, A.; Kamarianakis, Y.; Chrysoulakis, N. Estimating urban PM_{10} and $PM_{2.5}$ concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data. *Remote Sens. Environ.* **2016**, *172*, 148–164. [CrossRef]
55. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *J. Geophys. Res. Atmos.* **2009**, *114*. [CrossRef]
56. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **2016**, *50*, 4712–4721. [CrossRef]
57. Wu, Y.; Guo, J.; Zhang, X.; Li, X. Correlation between PM Concentrations and Aerosol Optical Depth in Eastern China Based on BP Neural Networks. In Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International, Vancouver, BC, Canada, 24–29 July 2011; pp. 3308–3311.
58. Grivas, G.; Chaloulakou, A. Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. *Atmos. Environ.* **2006**, *40*, 1216–1229. [CrossRef]
59. Hu, Z. Spatial analysis of MODIS aerosol optical depth, PM 2.5, and chronic coronary heart disease. *Int. J. Health Geogr.* **2009**, *8*, 27. [CrossRef] [PubMed]
60. Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. *Environ. Sci. Technol.* **2014**, *48*, 7436–7444. [CrossRef]
61. You, W.; Zang, Z.; Zhang, L.; Li, Y.; Pan, X.; Wang, W. National-scale estimates of ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. *Remote Sens.* **2016**, *8*, 184. [CrossRef]
62. Bai, Y.; Wu, L.; Qin, K.; Zhang, Y.; Shen, Y.; Zhou, Y. A geographically and temporally weighted regression model for ground-level $PM_{2.5}$ estimation from satellite-derived 500 m resolution AOD. *Remote Sens.* **2016**, *8*, 262. [CrossRef]
63. Paciorek, C.J.; Liu, Y.; Moreno-Macias, H.; Kondragunta, S. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level $PM_{2.5}$. *Environ. Sci. Technol.* **2008**, *42*, 5800–5806. [CrossRef]
64. Liu, Y.; Paciorek, C.J.; Koutrakis, P. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* **2009**, *117*, 886. [CrossRef]
65. Song, Y.-Z.; Yang, H.-L.; Peng, J.-H.; Song, Y.-R.; Sun, Q.; Li, Y. Estimating $PM_{2.5}$ concentrations in Xi'an City using a generalized additive model with multi-source monitoring data. *PLoS ONE* **2015**, *10*, e0142149. [CrossRef] [PubMed]
66. Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z. Satellite based mapping of ground $PM_{2.5}$ concentration using generalized additive modeling. *Remote Sens.* **2016**, *9*, 1. [CrossRef]
67. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes Jr, M.G.; Estes, S.M.; Quattrochi, D.A.; Puttaswamy, S.J. Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sens. Environ.* **2014**, *140*, 220–232. [CrossRef]
68. Ma, Z.; Hu, X.; Sayer, A.M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004–2013. *Environ. Health Perspect.* **2016**, *124*, 184. [CrossRef]
69. Li, L.; Chen, B.; Zhang, Y.; Zhao, Y.; Xian, Y.; Xu, G.; Zhang, H.; Guo, L. Retrieval of Daily $PM_{2.5}$ Concentrations Using Nonlinear Methods: A Case Study of the Beijing–Tianjin–Hebei Region, China. *Remote Sens.* **2018**, *10*, 2006. [CrossRef]
70. Danesh Yazdi, M.; Kuang, Z.; Dimakopoulou, K.; Barratt, B.; Suel, E.; Amini, H.; Lyapustin, A.; Katsouyanni, K.; Schwartz, J. Predicting Fine Particulate Matter ($PM_{2.5}$) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. *Remote Sens.* **2020**, *12*, 914. [CrossRef]
71. Kleine Deters, J.; Zalakeviciute, R.; Gonzalez, M.; Rybarczyk, Y. Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters. *J. Electr. Comput. Eng.* **2017**, *2017*. [CrossRef]
72. Shin, M.; Kang, Y.; Park, S.; Im, J.; Yoo, C.; Quackenbush, L.J. Estimating ground-level particulate matter concentrations using satellite-based data: A review. *GIScience Remote Sens.* **2020**, *57*, 174–189. [CrossRef]
73. Gholami, H.; Mohamadifar, A.; Sorooshian, A.; Jansen, J.D. Machine-learning algorithms for predicting land susceptibility to dust emissions: The case of the Jazmurian Basin, Iran. *Atmos. Pollut. Res.* **2020**, *11*, 1303–1315. [CrossRef]
74. Gholami, H.; Mohammadifar, A.; Bui, D.T.; Collins, A.L. Mapping wind erosion hazard with regression-based machine learning algorithms. *Sci. Rep.* **2020**, *10*, 1–16. [CrossRef] [PubMed]
75. Guo, Y.; Tang, Q.; Gong, D.-Y.; Zhang, Z. Estimating ground-level $PM_{2.5}$ concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sens. Environ.* **2017**, *198*, 140–149. [CrossRef]
76. Hu, X.; Waller, L.A.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes Jr, M.G.; Estes, S.M.; Quattrochi, D.A.; Sarnat, J.A.; Liu, Y. Estimating ground-level $PM_{2.5}$ concentrations in the southeastern US using geographically weighted regression. *Environ. Res.* **2013**, *121*, 1–10. [CrossRef]
77. Jiang, M.; Sun, W.; Yang, G.; Zhang, D. Modelling seasonal GWR of daily $PM_{2.5}$ with proper auxiliary variables for the Yangtze River Delta. *Remote Sens.* **2017**, *9*, 346. [CrossRef]
78. Luo, J.; Du, P.; Samat, A.; Xia, J.; Che, M.; Xue, Z. Spatiotemporal pattern of PM 2.5 concentrations in mainland China and analysis of its influencing factors using geographically weighted regression. *Sci. Rep.* **2017**, *7*, 1–14.
79. Gholami, H.; Kordestani, M.D.; Li, J.; Telfer, M.W.; Fathabadi, A. Diverse sources of aeolian sediment revealed in an arid landscape in southeastern Iran using a modified Bayesian un-mixing model. *Aeolian Res.* **2019**, *41*, 100547. [CrossRef]

80. Kusuma, W.L.; Chih-Da, W.; Yu-Ting, Z.; Hapsari, H.H.; Muhamad, J.L. $PM_{2.5}$ pollutant in Asia—a comparison of metropolis cities in Indonesia and Taiwan. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4924. [CrossRef] [PubMed]
81. Amil, N.; Latif, M.T.; Khan, M.F.; Mohamad, M. Seasonal variability of PM 2.5 composition and sources in the Klang Valley urban-industrial environment. *Atmos. Chem. Phys.* **2016**, *16*, 5357–5381. [CrossRef]
82. Grange, S.K.; Carslaw, D.C.; Lewis, A.C.; Boleti, E.; Hueglin, C. Random forest meteorological normalisation models for Swiss PM 10 trend analysis. *Atmos. Chem. Phys.* **2018**, *18*, 6223–6239. [CrossRef]
83. Lovrić, M.; Pavlović, K.; Vuković, M.; Grange, S.K.; Haberl, M.; Kern, R. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. *Environ. Pollut.* **2021**, *274*, 115900. [CrossRef]
84. Šimić, I.; Lovrić, M.; Godec, R.; Kröll, M.; Bešlić, I. Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon. *Environ. Pollut.* **2020**, *263*, 114587. [CrossRef]
85. Shaziayani, W.N.; Ul-Saufie, A.Z.; Libasin, Z.; Shukri, F.N.A.; Abdullah, S.S.S.; Noor, N.M. A Review of PM_{10} Concentrations Modelling in Malaysia. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Seoul, Korea, 23–24 July 2020; p. 12008.
86. Dahari, N.; Latif, M.T.; Muda, K.; Hussein, N. Influence of Meteorological Variables on Suburban Atmospheric $PM_{2.5}$ in the Southern Region of Peninsular Malaysia. *Aerosol Air Qual. Res.* **2020**, *20*, 14–25. [CrossRef]
87. Ee-Ling, O.; Mustaffa, N.I.H.; Amil, N.; Khan, M.F.; Latif, M.T. Source contribution of PM 2.5 at different locations on the Malaysian Peninsula. *Bull. Environ. Contam. Toxicol.* **2015**, *94*, 537–542. [CrossRef] [PubMed]
88. Yaakob, U.; Masron, T.; Masami, F. Ninety years of urbanization in Malaysia: A geographical investigation of its trends and characteristics. *J. Ritsumeikan Soc. Sci Hum.* **2010**, *4*, 79–101.
89. Jamil, A.; Makmom, A.A.; Saeid, P.; Firuz, R.M.; Prinaz, R. PM_{10} monitoring using MODIS AOT and GIS, Kuala Lumpur, Malaysia. *Res. J. Chem. Environ.* **2011**, *15*, 2.
90. Abas, M.R.; Oros, D.R.; Simoneit, B.R. Biomass burning as the main source of organic aerosol particulate matter in Malaysia during haze episodes. *Chemosphere* **2004**, *55*, 1089–1095. [CrossRef] [PubMed]
91. Awang, M.B.; Jaafar, A.B.; Abdullah, A.M.; Ismail, M.B.; Hassan, M.N.; Abdullah, R.; Johan, S.; Noor, H. Air quality in Malaysia: Impacts, management issues and future challenges. *Respirology* **2000**, *5*, 183–196. [CrossRef]
92. Kanniah, K.D.; Zaman, N.A.F.K.; Kaskaoutis, D.G.; Latif, M.T. COVID-19's impact on the atmospheric environment in the Southeast Asia region. *Sci. Total Environ.* **2020**, *736*, 139658. [CrossRef] [PubMed]
93. Ash'aari, Z.H.; Aris, A.Z.; Ezani, E.; Kamal, N.I.A.; Jaafar, N.; Jahaya, J.N.; Manan, S.A.; Saifuddin, M.F.U. Spatiotemporal variations and contributing factors of air pollutant concentrations in Malaysia during movement control order due to pandemic COVID-19. *Aerosol Air Qual. Res.* **2020**, *20*, 2047–2061. [CrossRef]
94. Alhasa, K.M.; Mohd Nadzir, M.S.; Olalekan, P.; Latif, M.T.; Yusup, Y.; Iqbal Faruque, M.R.; Ahamad, F.; Aiyub, K.; Md Ali, S.H.; Khan, M.F. Calibration model of a low-cost air quality sensor using an adaptive neuro-fuzzy inference system. *Sensors* **2018**, *18*, 4380. [CrossRef]
95. Latif, M.T.; Dominick, D.; Hawari, N.S.S.L.; Mohtar, A.A.A.; Othman, M. The concentration of major air pollutants during the movement control order due to the COVID-19 pandemic in the Klang Valley, Malaysia. *Sustain. Cities Soc.* **2021**, *66*, 102660. [CrossRef]
96. Li, T.; Zhang, C.; Shen, H.; Yuan, Q.; Zhang, L. Real-time and Seamless Monitoring of Ground-Level $PM_{2.5}$ Using Satellite Remote Sensing. *arXiv* **2018**, arXiv:1803.03409.
97. Sowden, M.; Mueller, U.; Blake, D. Review of surface particulate monitoring of dust events using geostationary satellite remote sensing. *Atmos. Environ.* **2018**, *183*, 154–164. [CrossRef]
98. Liu, J.; Weng, F.; Li, Z.; Cribb, M.C. Hourly $PM_{2.5}$ Estimates from a Geostationary Satellite Based on an Ensemble Learning Algorithm and Their Spatiotemporal Patterns over Central East China. *Remote Sens.* **2019**, *11*, 2120. [CrossRef]
99. Zang, L.; Mao, F.; Guo, J.; Gong, W.; Wang, W.; Pan, Z. Estimating hourly PM1 concentrations from Himawari-8 aerosol optical depth in China. *Environ. Pollut.* **2018**, *241*, 654–663. [CrossRef] [PubMed]
100. She, L.; Xue, Y.; Yang, X.; Guang, J.; Li, Y.; Che, Y.; Fan, C.; Xie, Y. Dust detection and intensity estimation using Himawari-8/AHI observation. *Remote Sens.* **2018**, *10*, 490. [CrossRef]
101. Yumimoto, K.; Nagao, T.; Kikuchi, M.; Sekiyama, T.; Murakami, H.; Tanaka, T.; Ogi, A.; Irie, H.; Khatri, P.; Okumura, H. Aerosol data assimilation using data from Himawari-8, a next-generation geostationary meteorological satellite. *Geophys. Res. Lett.* **2016**, *43*, 5886–5894. [CrossRef]
102. Kikuchi, M.; Murakami, H.; Suzuki, K.; Nagao, T.M.; Higurashi, A. Improved hourly estimates of aerosol optical thickness using spatiotemporal variability derived from Himawari-8 geostationary satellite. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3442–3455. [CrossRef]
103. Chen, J.; Huang, X. Estimating Hourly PM 2.5 Concentrations from Himawari-8 AOD over Hubei Province. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*. [CrossRef]
104. Zaman, N.A.F.K.; Kanniah, K.D. Spatio-temporal assessment of Aerosol Optical Depth from Himawari-8 Satellite Data over Malaysia. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Kuala Lumpur, Malaysia, 20–21 October 2020; p. 12053.
105. Wei, H.; Wang, W.; Xu, F.; Feng, J. Evaluation of the Himawari-8 Aerosol Products. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 7825–7828.

106. Choi, M.; Lim, H.; Kim, J.; Lee, S.; Eck, T.F.; Holben, B.N.; Garay, M.J.; Hyer, E.J.; Saide, P.E.; Liu, H. Validation, comparison, and integration of GOCI, AHI, MODIS, MISR, and VIIRS aerosol optical depth over East Asia during the 2016 KORUS-AQ campaign. *Meas. Tech.* **2019**, *12*, 4619–4641. [CrossRef]
107. Wei, J.; Li, Z.; Sun, L.; Peng, Y.; Zhang, Z.; Li, Z.; Su, T.; Feng, L.; Cai, Z.; Wu, H. Evaluation and uncertainty estimate of next-generation geostationary meteorological Himawari-8/AHI aerosol products. *Sci. Total Environ.* **2019**, *692*, 879–891. [CrossRef] [PubMed]
108. Liu, B.-C.; Binaykia, A.; Chang, P.-C.; Tiwari, M.K.; Tsao, C.-C. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLoS ONE* **2017**, *12*, e0179763. [CrossRef] [PubMed]
109. Wu, C.; Wu, D.; Yu, J.Z. Estimation and uncertainty analysis of secondary organic carbon using 1 year of hourly organic and elemental carbon data. *J. Geophys. Res. Atmos.* **2019**, *124*, 2774–2795. [CrossRef]
110. Su, T.; Li, Z.; Li, C.; Li, J.; Han, W.; Shen, C.; Tan, W.; Wei, J.; Guo, J. The significant impact of aerosol vertical structure on lower atmosphere stability and its critical role in aerosol–planetary boundary layer (PBL) interactions. *Atmos. Chem. Phys.* **2020**, *20*, 3713–3724. [CrossRef]
111. Zang, L.; Wang, Z.; Zhu, B.; Zhang, Y. Roles of relative humidity in aerosol pollution aggravation over Central China during wintertime. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4422. [CrossRef]
112. Wang, J.; Martin, S.T. Satellite characterization of urban aerosols: Importance of including hygroscopicity and mixing state in the retrieval algorithms. *J. Geophys. Res. Atmos.* **2007**, *112*. [CrossRef]
113. Titos, G.; Cazorla, A.; Zieger, P.; Andrews, E.; Lyamani, H.; Granados-Muñoz, M.J.; Olmo, F.; Alados-Arboledas, L. Effect of hygroscopic growth on the aerosol light-scattering coefficient: A review of measurements, techniques and error sources. *Atmos. Environ.* **2016**, *141*, 494–507. [CrossRef]
114. Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.; Hamid, H.A. Performance of multiple linear regression model for long-term PM_{10} concentration prediction based on gaseous and meteorological parameters. *J. Appl. Sci.* **2012**, *12*, 1488–1494. [CrossRef]
115. Abdullah, S.; Ismail, M.; Ahmed, A.N.; Abdullah, A.M. Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. *Atmosphere* **2019**, *10*, 667. [CrossRef]
116. Ahmad, M.; Alam, K.; Tariq, S.; Anwar, S.; Nasir, J.; Mansha, M. Estimating fine particulate concentration using a combined approach of linear regression and artificial neural network. *Atmos. Environ.* **2019**, *219*, 117050. [CrossRef]
117. Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; De Hoogh, K.; De'Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M. Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170–179. [CrossRef]
118. Mhawish, A.; Banerjee, T.; Sorek-Hamer, M.; Bilal, M.; Lyapustin, A.I.; Chatfield, R.; Broday, D.M. Estimation of high-resolution $PM_{2.5}$ over the indo-gangetic plain by fusion of satellite data, meteorology, and land use variables. *Environ. Sci. Technol.* **2020**, *54*, 7891–7900. [CrossRef]
119. Jiang, T.; Chen, B.; Nie, Z.; Ren, Z.; Xu, B.; Tang, S. Estimation of hourly full-coverage $PM_{2.5}$ concentrations at 1-km resolution in China using a two-stage random forest model. *Atmos. Res.* **2021**, *248*, 105146. [CrossRef]
120. Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, USA, 2015; pp. 67–80. [CrossRef]
121. Weizhen, H.; Zhengqiang, L.; Yuhuan, Z.; Hua, X.; Ying, Z.; Kaitao, L.; Donghui, L.; Peng, W.; Yan, M. Using Support Vector Regression to Predict PM_{10} and $PM_{2.5}$. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Beijing, China, 22–26 April 2013; p. 12268.
122. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
123. Sage, A. Random Forest Robustness, Variable Importance, and Tree Aggregation. 2018. Available online: https://lib.dr.iastate.edu/etd/16453 (accessed on 20 May 2021).
124. Hou, N.; Zhang, X.; Zhang, W.; Wei, Y.; Jia, K.; Yao, Y.; Jiang, B.; Cheng, J. Estimation of Surface Downward Shortwave Radiation over China from Himawari-8 AHI Data Based on Random Forest. *Remote Sens.* **2020**, *12*, 181. [CrossRef]
125. Strobl, C.; Malley, J.; Tutz, G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **2009**, *14*, 323. [CrossRef]
126. Janitza, S.; Hornung, R. On the overestimation of random forest's out-of-bag error. *PLoS ONE* **2018**, *13*, e0201904. [CrossRef]
127. 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. Technol.* **2017**, *51*, 6936–6944. [CrossRef] [PubMed]
128. Brokamp, C.; Jandarov, R.; Hossain, M.; Ryan, P. Predicting daily urban fine particulate matter concentrations using a random forest model. *Environ. Sci. Technol.* **2018**, *52*, 4173–4179. [CrossRef]
129. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
130. Li, T.; Shen, H.; Zeng, C.; Yuan, Q. A Validation approach considering the uneven distribution of ground stations for satellite-based PM 2.5 estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1312–1321. [CrossRef]
131. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 1–21. [CrossRef]

132. Kalita, G.; Kunchala, R.K.; Fadnavis, S.; Kaskaoutis, D.G. Long term variability of carbonaceous aerosols over Southeast Asia via reanalysis: Association with changes in vegetation cover and biomass burning. *Atmos. Res.* **2020**, *245*, 105064. [CrossRef]
133. Pani, S.K.; Lin, N.-H.; Griffith, S.M.; Chantara, S.; Lee, C.-T.; Thepnuan, D.; Tsai, Y.I. Brown carbon light absorption over an urban environment in northern peninsular Southeast Asia. *Environ. Pollut.* **2021**, *276*, 116735. [CrossRef]
134. Nguyen, T.P.M.; Bui, T.H.; Nguyen, M.K.; Nguyen, T.H.; Pham, H.L. Impact of COVID-19 partial lockdown on PM 2.5, SO_2, NO_2, O_3, and trace elements in PM 2.5 in Hanoi, Vietnam. *Environ. Sci. Pollut. Res.* **2021**, 1–11. [CrossRef]
135. Grivas, G.; Dimakopoulou, K.; Samoli, E.; Papakosta, D.; Karakatsani, A.; Katsouyanni, K.; Chaloulakou, A. Ozone exposure assessment for children in Greece-Results from the RESPOZE study. *Sci. Total Environ.* **2017**, *581*, 518–529. [CrossRef]
136. Hatzianastassiou, N.; Katsoulis, B.D.; Antakis, B. Extreme nitrogen oxide and ozone concentrations in Athens atmosphere in relation to meteorological conditions. *Environ. Monit. Assess.* **2007**, *128*, 447–464. [CrossRef] [PubMed]
137. Jin, Q.; Crippa, P.; Pryor, S. Spatial characteristics and temporal evolution of the relationship between $PM_{2.5}$ and aerosol optical depth over the eastern USA during 2003–2017. *Atmos. Environ.* **2020**, *239*, 117718. [CrossRef]
138. Gratsea, M.; Liakakou, E.; Mihalopoulos, N.; Adamopoulos, A.; Tsilibari, E.; Gerasopoulos, E. The combined effect of reduced fossil fuel consumption and increasing biomass combustion on Athens' air quality, as inferred from long term CO measurements. *Sci. Total Environ.* **2017**, *592*, 115–123. [CrossRef]
139. Chelani, A.B. Estimating $PM_{2.5}$ concentration from satellite derived aerosol optical depth and meteorological variables using a combination model. *Atmos. Pollut. Res.* **2019**, *10*, 847–857. [CrossRef]
140. Saraswat, I.; Mishra, R.K.; Kumar, A. Estimation of PM_{10} concentration from Landsat 8 OLI satellite imagery over Delhi, India. *Remote Sens. Appl. Soc. Environ.* **2017**, *8*, 251–257. [CrossRef]
141. Li, X.; Zhang, X. Predicting ground-level $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach. *Environ. Pollut.* **2019**, *249*, 735–749. [CrossRef]
142. Stavroulas, I.; Bougiatioti, A.; Grivas, G.; Paraskevopoulou, D.; Tsagkaraki, M.; Zarmpas, P.; Liakakou, E.; Gerasopoulos, E.; Mihalopoulos, N. Sources and processes that control the submicron organic aerosol composition in an urban Mediterranean environment (Athens): A high temporal-resolution chemical composition measurement study. *Atmos. Chem. Phys.* **2019**, *19*, 901–919. [CrossRef]
143. Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophys. Res. Lett.* **2017**, *44*, 11985–11993. [CrossRef]
144. Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; 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]
145. Park, S.; Lee, J.; Im, J.; Song, C.-K.; Choi, M.; Kim, J.; Lee, S.; Park, R.; Kim, S.-M.; Yoon, J. 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.* **2020**, *713*, 136516. [CrossRef]
146. Li, T.; Shen, H.; Zeng, C.; Yuan, Q.; Zhang, L. Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: Methods and assessment. *Atmos. Environ.* **2017**, *152*, 477–489. [CrossRef]
147. Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M.L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* **2017**, *155*, 129–139. [CrossRef]
148. Xu, Y.; Ho, H.C.; Wong, M.S.; Deng, C.; Shi, Y.; Chan, T.-C.; Knudby, A. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$. *Environ. Pollut.* **2018**, *242*, 1417–1426. [CrossRef]
149. Shen, H.; Zhou, M.; Li, T.; Zeng, C. Integration of remote sensing and social sensing data in a deep learning framework for hourly urban $PM_{2.5}$ mapping. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4102. [CrossRef] [PubMed]
150. Chen, W.; Ran, H.; Cao, X.; Wang, J.; Teng, D.; Chen, 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]
151. Dutta, A.; Jinsart, W. Air Pollution in Indian Cities and Comparison of MLR, ANN and CART Models for Predicting PM_{10} Concentrations in Guwahati, India. *Asian J. Atmos. Environ.* **2021**, *15*, 2020131. [CrossRef]
152. Deng, X.; Tie, X.; Zhou, X.; Wu, D.; Zhong, L.; Tan, H.; Li, F.; Huang, X.; Bi, X.; Deng, T. Effects of Southeast Asia biomass burning on aerosols and ozone concentrations over the Pearl River Delta (PRD) region. *Atmos. Environ.* **2008**, *42*, 8493–8501. [CrossRef]
153. Sinha, P.; Gupta, P.; Kaskaoutis, D.; Sahu, L.; Nagendra, N.; Manchanda, R.; Kumar, Y.B.; Sreenivasan, S. Estimation of particulate matter from satellite-and ground-based observations over Hyderabad, India. *Int. J. Remote Sens.* **2015**, *36*, 6192–6213. [CrossRef]
154. Wang, X.; Sun, W. Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous $PM_{2.5}$ concentrations using deep neural network in Beijing–Tianjin–Hebei, China. *Atmos. Environ.* **2019**, *211*, 128–137. [CrossRef]
155. Ibrahim, M.Z.; Ismail, M.; Yong, K.H. Mapping the Spatial Distribution of Criteria Air Pollutants in Peninsular Malaysia Using Geographical Information System (GIS). *Tech. Air Pollut. Monit. Model. Health* **2012**, 153. [CrossRef]
156. Van Donkelaar, A.; Martin, R.V.; Brauer, M.; Kahn, R.; Levy, R.; Verduzco, C.; Villeneuve, P.J. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environ. Health Perspect.* **2010**, *118*, 847. [CrossRef]