remote sensing

MDPI

Article

Near-Surface NO_2 Concentration Estimation by Random Forest Modeling and Sentinel-5P and Ancillary Data

Meixin Li [1,2], Ying Wu [1,3,*], Yansong Bao [1,3], Bofan Liu [1,4] and George P. Petropoulos [5]

[1] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Key Laboratory for Aerosol-Cloud-Precipitation of China Meteorological Administration, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20181204011@nuist.edu.cn (M.L.); ysbao@nuist.edu.cn (Y.B.); bofan.liu@nuist.edu.cn (B.L.)

[2] Focused Photonics (Hangzhou) Inc., Hangzhou 310052, China

[3] School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing 210044, China

[4] Reading Academy, Nanjing University of Information Science and Technology, Nanjing 210044, China

[5] Department of Geography, Harokopio University of Athens, 17671 Athens, Greece; gpetropoulos@hua.gr

* Correspondence: yingwu@nuist.edu.cn

Abstract: In the present study, a daily model is proposed for estimating the near-surface NO_2 concentration in China, combining for the first time the Random Forest (RF) machine learning algorithm with the tropospheric NO_2 columns from the TROPOspheric Monitoring Instrument (TropOMI) satellite and meteorological and NO_2 data of surface sites in China for the year 2019. Furthermore, near-surface NO_2 concentration data of ground sites during the COVID-19 outbreak from 1–5 February 2020 were used to verify the developed model. The daily model was verified by the ten-fold cross-validation method, revealing a coefficient of determination (R^2) of 0.78 and root-mean-square error (RMSE) of 7.04 $\mu g/m^3$, which are reasonable and also comparable to other published studies. In addition, our model showed that near-surface NO_2 in China during the COVID-19 pandemic was significantly reduced compared with 2019, and these predictions were in good agreement with reference ground data. Our proposed model can also provide NO_2 estimates for areas in western China where there are few ground monitoring sites. Therefore, all in all, our study findings suggest that the model established herein is suitable for estimating the daily NO_2 concentration near the surface in China and, as such, can be used if there is a lack of surface sites and/or missing observations in some areas.

Keywords: Random Forest; NO_2; TropOMI; near-surface pollution; remote sensing

check for updates

Citation: Li, M.; Wu, Y.; Bao, Y.; Liu, B.; Petropoulos, G.P. Near-Surface NO_2 Concentration Estimation by Random Forest Modeling and Sentinel-5P and Ancillary Data. *Remote Sens.* **2022**, *14*, 3612. https://doi.org/10.3390/rs14153612

Academic Editor: Ka Lok Chan

Received: 22 June 2022
Accepted: 25 July 2022
Published: 28 July 2022

1. Introduction

Nitrogen oxides (NO_x) are an important trace gas in the atmosphere for a variety of atmospheric pollutants, and all forms of combustion result in the release of NO_x [1,2]. They have a short life span (existing in the atmosphere for 1–12 h) [3], but they play a key role in the process of photochemically induced ozone (O_3) catalysis, which leads to summertime smog and elevated levels of global tropospheric O_3 [4,5]. The two main types of NO_x are nitric oxide (NO) and nitrogen dioxide (NO_2) [6]. The large amounts of NO_x emissions in China have resulted in high concentrations and a high increase rate of tropospheric NO_2—much higher than those in other countries and regions [7]. As an air pollutant, NO_2 is a brownish-red toxic and pungent gas at high temperatures and is an important indicator of air quality at ground level [8]. At high concentrations it causes adverse effects on human health, climate change and the ecological environment [9–11].

In recent years, due to increasingly severe environmental pollution, compared with the tropospheric NO_2 concentration monitored by satellites, the near-surface NO_2 concentration is more closely related to human life and the ecological environment and has therefore been a key area of study. In this respect, due to the specific regional limitations related to the monitoring of NO_2 concentrations at ground sites, research on the spatial distribution of near-surface NO_2 concentrations has attracted more and more attention. Exploiting tropospheric NO_2 column concentration data available from satellite products, a few research studies have confirmed that there is a certain correlation between tropospheric NO_2 column concentrations and near-ground-monitored NO_2 concentrations [12–14]. On this basis, Lamsal et al. [15] estimated the near-surface NO_2 concentration in North America using the local scale factor obtained from a global three-dimensional model (GEOS-Chem) and the tropospheric NO_2 column concentration retrieved by the Ozone Monitoring Instrument (OMI) [16,17] on board the Aura satellite platform. The authors reported that the difference between winter and spring was higher than that between summer and autumn, which proved there was a seasonal difference in NO_2 concentration. The authors reported a correlation coefficient (CC) of 0.86 with even higher values in heavily polluted areas.

Considering the continuous reduction in the number of ground-based NO_2 concentration monitoring sites, satellite remote sensing may become the most effective method to monitor NO_2 concentrations. It is also found that the NO_2 concentration in the boundary layer may play a major role in the tropospheric NO_2 column concentration in polluted areas [18,19]. For instance, the tropospheric NO_2 column concentration retrieved by the Global Ozone Monitoring Experiment (GOME) [20] and Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) [21] was found to be closely related to quantities of NO_x emissions at the land surface [22,23]. Some researchers have also reported an important relationship between the near-ground monitoring of NO_2 concentrations and GOME's tropospheric NO_2 column concentration [24,25]. Qin et al. [26] evaluated the ability of four methods to estimate the near-surface NO_2 concentration in eastern China through OMI's tropospheric NO_2 column concentration data—namely, the least-squares method, geographically weighted regression, time-weighted regression, and geographic time-weighted regression. The authors reported the geographic time-weighted regression method as the most promising one. Some researchers have also worked on the assumption that there is a close relationship between near-surface NO_2 concentrations and the tropospheric NO_2 column monitored by satellites, which may be due to two reasons. One is that human activities are mainly concentrated near the surface, so the ground-level NO_2 column concentration is the main source of the tropospheric NO_2 column concentration. The other reason is the short lifetime of near-surface NO_2, resulting in little transmission in both vertical and horizontal directions [24,27]. Bechle et al. [28] used the tropospheric NO_2 column concentration product monitored by the OMI to further estimate the near-surface NO_2 concentration on the basis of scaling factors (surface-to-column ratios). Comparing the results with ground-monitored NO_2 data, it was found that the annual average deviation was 13% and the seasonal deviation was between 1% and 22%. In another study, Gu et al. [29] combined the whole-layer NO_2 column concentration product monitored by the OMI with the NO_2 profile simulated by the Community Multi-scale Air Quality (CMAQ) [30] model to obtain the NO_2 concentration near the ground, and compared the results with the ground-monitored data of China's National Environmental Monitoring Center (CNEMC). A correlation of 0.75 between the CMAQ model alone and ground measurements was found, which increased to 0.80 and 0.78 for January and July, respectively, after adding satellite data. Based on data such as TropOMI TCD NO_2, Ghahremanloo et al. [31] used Deep Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Random Forest (RF) and Multiple Linear Regression (MLR) methods to obtain a model of daily near-ground NO_2 concentration in Texas, USA, and finally concluded that the Deep CNN model estimated the best results, with a correlation coefficient (R) of 0.91, an index of agreement (IOA) of 0.95, and a mean absolute bias (MAB) of 1.75 ppb. Chen et al. [32] compared the ability of three different methods to estimate the daily NO_2 across mainland China during 2014–2016, and it turned out that the kriging-calibrated satellite method was the best, with model R-squared and root mean square error (RMSE) of 0.85 and 7.87 $\mu g/m^3$. Lee et al. [33] estimated daily ground-level NO_2 concentrations in

New England, USA for the period 2005–2010 using a space-bound land-use regression of NO_2 density in the tropospheric column from the OMI. A mixed effects model showed a reasonably high predictive power for daily NO_2 concentrations (cross-validation R2 = 0.79).

In purview of the above, the present study aims at: (1) establishing a daily model suitable for estimating near-surface NO_2 concentrations in China, combining for the first time the RF machine learning (ML) method and the tropospheric NO_2 concentration product of the high-resolution TROPOspheric Monitoring Instrument (TropOMI). The proposed method also utilizes meteorological forecast data from the Global Forecast System of the National Centers for Environmental Prediction (NCEP/GFS), normalized difference vegetation index (NDVI) data from the Moderate Resolution Imaging Spectroradiometer (MODIS), and date, longitude, latitude and other auxiliary data of the same period; (2) comparing predictions of the proposed model with the monitored values at ground sites during the COVID-19 outbreak in 2020 to verify its accuracy and timeliness. The model developed herein is suitable for estimating daily concentrations of near-surface NO_2 in China and can successfully address issues arising from the limited number of ground sites and missing observations in some areas.

2. Dataset Construction

2.1. Data Preparation

2.1.1. TropOMI Data

This study uses the TropOMI offline version of the tropospheric NO_2 column concentration L2 product based on the inversion of differential absorption spectroscopy released by the European Space Agency (ESA). TropOMI is a tropospheric observer carried on board the Sentinel-5P satellite launched by the European Space Agency (ESA) on 13 October 2017. This satellite is a solar synchronous orbit satellite operating at an altitude of 824 km from the ground, with an operation cycle of 17 days [34]. Since 6 August 2019, TropOMI has had a spatial resolution of 7 km × 3.5 km, improved from the 5.5 km × 3.5 km prior to that date. The latter allowed monitoring the composition of various trace gases in the global atmosphere effectively, thereby strengthening the monitoring of clouds and aerosols [35,36]. TropOMI has two kinds of tropospheric NO_2 products. One is a near-real-time product, which allows users to obtain the imaging data of an area within 4 h after the satellite has scanned that area. This option facilitates researchers to understand the spatial distribution characteristics of pollutant concentrations in a timely manner and further study and analyze the pollutant situation in an area [37,38]. The other is an offline product that allows satellite data to be downloaded from the official website within a few days.

A comparison of the main parameters between the tropospheric NO_2 column concentration L2 product of TropOMI and other satellite sensor products is provided in Table 1. As can be seen, the TropOMI NO_2 column concentration L2 product represents an improvement over the inversion method of the OMI L2 product; it also improves the wavelength calibration of the L1 data spectrum. In addition, the fitting result of the NO_2 column concentration can be improved by adding equal trace gases into the fitting. TropOMI confirms the best fitting based on ensuring that the error and root-mean-square error (RMSE) of the NO_2 inclined column are reduced and the auxiliary trace gas that significantly improves the degree of fitting is included [39]. Subsequently, improvements have been made in stratospheric tropospheric separation and the calculation of atmospheric quality factors [40].

In this study, the part of the TropOMI tropospheric NO_2 column concentration dataset in which the quality of the data exceeds a certain threshold [(qa_values) > 0.75] is selected, which excludes those pixels in which snow/ice-covered areas were observed or the data include incorrect values.

Table 1. NO_2 inversion algorithm of TropOMI compared with previous sensors(Adapted with permission from Ref. [41]. 2017, Gu, J.; Chen, L.; Yu, C.; Li, S.; Tao, J.; Fan, M.; Xiong, X.; Wang, Z.; Shang, H.; Su, L.).

	TropOMI	**OMI**	**GOME-2**	**SCIAMACHY**
Wavelength range (nm)	405–465	405–465	425–450	426.5–451.5
Secondary trace gases	O_3, H_2O_{vap}, O_2-O_2, H_2O_{lip}	O_3, H_2O_{vap}, O_2-O_2, H_2O_{lip}	O_3, H_2O_{vap}, O_2-O_2	O_3, H_2O_{vap}, O_2-O_2
False absorber	Ring	Ring	Ring	Ring
Fitting method	Non-linear	Non-linear	Linear	Linear
Polynomial term number	5	5	3	2
Polarization calibration	No	No	No	Yes

2.1.2. Meteorological Data

The RF machine learning model established in this study uses meteorological data in 2019. Those include the temperature at 2 m height, the humidity at 1000 hPa, the zonal (U) and meridional (V) components of the wind, the atmospheric boundary layer height, and NDVI (shown in Table 2). Among them, the temperature, relative humidity, U and V wind components, and the height of the atmospheric boundary layer are from the reanalysis grid data of NCEP/GFS. Since they need to match the transit time of TropOMI in China (12:00–15:00 daily), the forecast meteorological data used are $0.25^\circ \times 0.25^\circ$ grid data obtained by integrating 6 h of daily GFS data.

Table 2. Meteorological parameters used in modeling.

Name	**Unit**	**Abbreviation**
Temperature at 2 m height	°C	T
Relative humidity at 2 m height	%	RH
U component of wind at 2 m height	m/s	U_{grd}
V component of wind at 2 m height	m/s	V_{grd}
Boundary layer height	m	PBLH
Normalized vegetation index	-	NDVI

The NDVI data are from the spectral reflectance data of MODIS carried on board the Aqua satellite. As an afternoon detection satellite, Aqua flies from north to south and passes over the equator at around 13:30 local time every day. MODIS has a total of 36 spectral channels and the scanning spatial resolutions are mainly 250 m, 500 m and 1000 m. The primary data for calculating the NDVI are from the satellite data with 1 km resolution corrected by the MODIS sensor, and the NDVI calculation is as follows:

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}}, \quad (1)$$

where ρ_{NIR} is the reflectivity in the near infrared band and ρ_{Red} is the reflectivity of the red light band.

2.1.3. Ground Monitoring Station Data

The ground-monitored NO_2 concentration data used in this study are from the hourly data provided by CNEMC air quality real-time release platform. The research period is from 1 January to 31 December 2019. In this year, there were 1604 ground sites in China from January to May. After May, a number of ground monitoring sites were newly built, and the number of ground monitoring sites from June to December reached 1641.

2.2. Ground and Satellite Data Quality Control

Quality control of both ground and satellite data is required before modeling. For the ground data, the sites with CCs higher than 0.15 were ultimately selected. Moreover, errors higher than three times the standard deviation were omitted. For the satellite data the part of the TropOMI tropospheric NO_2 column concentration dataset in which the quality of the data exceeds a certain threshold, i.e., qa_values > 0.75, was selected.

2.2.1. Ground Data Quality Control

This study takes the NO_2 concentration data monitored by ground sites as the target value, i.e., the near-surface NO_2 concentration, so the reliability of the ground site data is particularly important. Therefore, before modeling, it is necessary to screen the ground site data in China in 2019 and select the ground sites most suitable for modeling.

According to the data of ground monitoring sites in 2019, there are 1605 sites in China from January to May and 1644 sites from June to December. Therefore, in accordance with the provisions of Environmental Air Quality Indicators for monitoring data, the site results in China for every day in 2019 were screened and 1641 ground sites were ultimately determined for calculation. Since TropOMI passes through China at around 13:30 (Beijing time) every day, the hourly data of the ground sites from 12:00 to 15:00 were selected to calculate the average concentration of the NO_2 concentration at a particular site on that day to match with the satellite products. The matching method for geographical location involved matching the satellite data at each ground site by using the inverse distance weighting interpolation method within a 0.1° longitude and latitude range of the ground site. After the quality control and matching of the ground monitoring data, the correlation between the daily concentration results of various sites in China in 2019 and the daily tropospheric NO_2 products retrieved by satellite was calculated. Since new sites were added after May, the number of days for some sites is less than one year. The distribution of CCs for 1641 sites in China in 2019 is shown in Figure 1, in which the ordinate represents the value of CC and the abscissa represents the station serial number. It can be seen that the correlation is a mix of high and low values, with the CC of some sites reaching close to 1, but others showing basically no correlation. All figures in this study are made using the visualization package in Python [42–44].

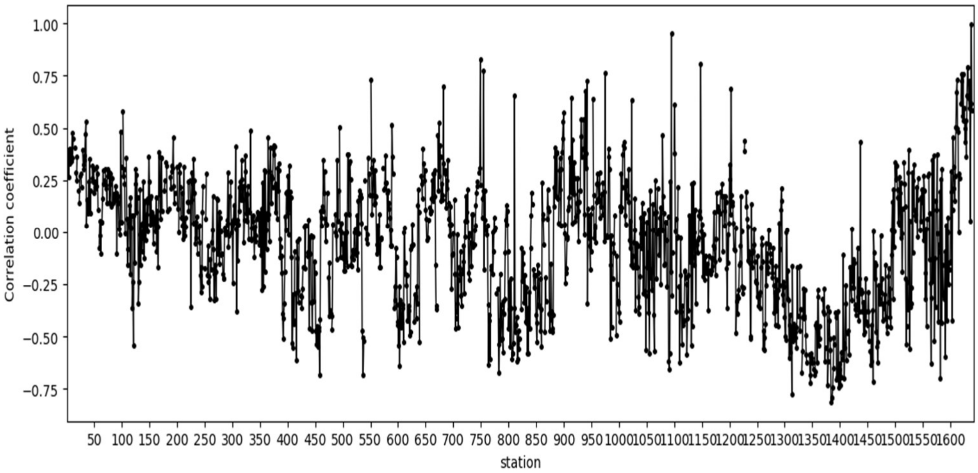


Figure 1. Correlation coefficient (CC) distribution of ground daily NO_2 concentration results and TropOMI products for 1641 sites in China in 2019.

The statistical results of the correlation distribution presented in Figure 1 are detailed in Table 3. The number of sites with CCs in different intervals were counted, and the CCs were divided into five grades, as follows: (i) the absolute value of the CC (|r|) is greater than 0.7; (ii) |r| is higher than 0.5 and less than 0.7; (iii) |r| is higher than 0.15 and less than 0.5; (iv) |r| is less than 0.15; and (v) the CC cannot be calculated because of missing data throughout the year. According to these results, it can be seen that most sites in China are concentrated in the CC range of higher than 0.15 and less than 0.5, and less than 0.15. There are 28 sites with CCs higher than 0.7, while there are 133 sites without correlation.

Table 3. Distribution of correlation coefficient (CC) values among 1641 sites.

Range	**$|r| > 0.7$**	**$0.5 \leq |r| < 0.7$**	**$0.15 \leq |r| < 0.5$**	**$|r| < 0.15$**	**Nan**
Number of sites	28	134	801	545	133

The spatial distribution of different CCs in China is illustrated in Figure 2a–d, which shows the CCs higher than 0.7, higher than 0.5 and less than 0.7, higher than 0.15 and less than 0.5, and less than 0.15, respectively. There are 28 ground sites with CCs higher than 0.7, mainly distributed in Central China (Box1) and Yunnan (Box2), and relatively few in areas with serious near-surface NO_2 pollution. The sites with CCs higher than 0.5 and less than 0.7 are basically distributed in Southwest China (Box3), where the pollution (particularly in Sichuan and Chongqing) is also relatively serious. However, there are many sites in the latter two grades (i.e., higher than 0.15 and less than 0.5, and less than 0.15), and they are randomly distributed throughout China and involve all cities. Therefore, according to the results shown in Figure 2 and Table 3, the sites with CCs higher than 0.15 were ultimately selected as the research object for modeling.

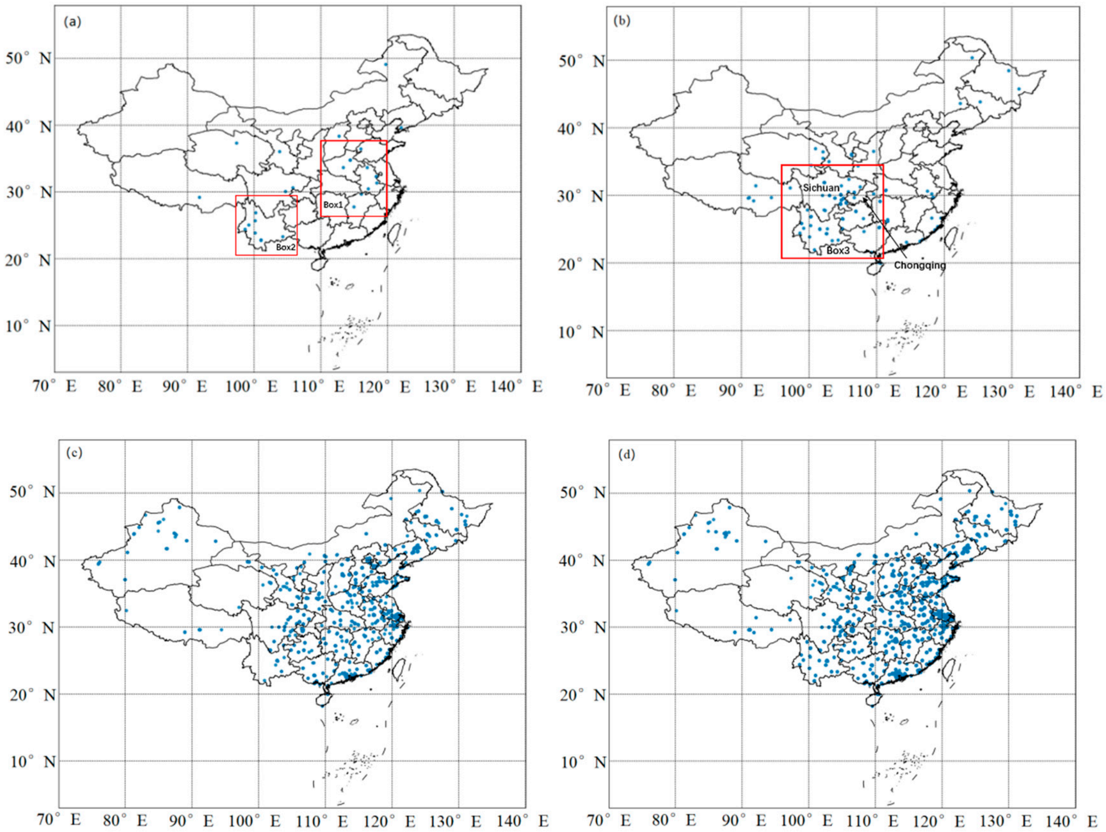


Figure 2. Site distribution with different correlation coefficient (CC) value ranges in China (red Box1, Central China; red Box2, Yunnan; red Box3, Southwest China): (**a**) $|r| > 0.7$; (**b**) $0.5 \leq |r| < 0.7$; (**c**) $0.15 \leq |r| < 0.5$; (**d**) $|r| < 0.15$.

In 2019, a total of 963 ground monitoring sites were selected in China as the target value for modeling, and several of them were randomly selected to study the NO_2 concentration distribution of these ground sites, all of which were subject to quality control according to the distribution. Figure 3 shows a comparison before and after quality control of three sites chosen randomly in China. From the distribution of the annual concentration of sites before quality control on the left, it can be seen that the first and second sites, which are the ground data sites, have days when the near-ground NO_2 pollution concentration is considerably higher than on other days. In particular, in the first site, there is a significant difference between the abnormal large value of individual days and the concentration of adjacent days. The third site has multiple days with serious pollution at the same time, inferring that there may be a trend of more serious pollution in this area. In view of the above, errors higher than three times the standard deviation were omitted to eliminate anomalies from instrument failure or human error, which may otherwise have affected the modeling data quality [45]. Specifically, values were eliminated by calculating three times the standard deviation of the NO_2 concentration of the site in one year, and the three panels on the right show the results after elimination. It can be seen that the site concentration after quality control is much smoother and there are days with very high NO_2 pollution. Therefore, this quality control method does not eliminate all days with serious pollution, which also hinders the process of machine learning. Therefore, herein, the data of all of the 963 ground sites in China in 2019 that reached this error threshold were eliminated to make the model target value more accurate.

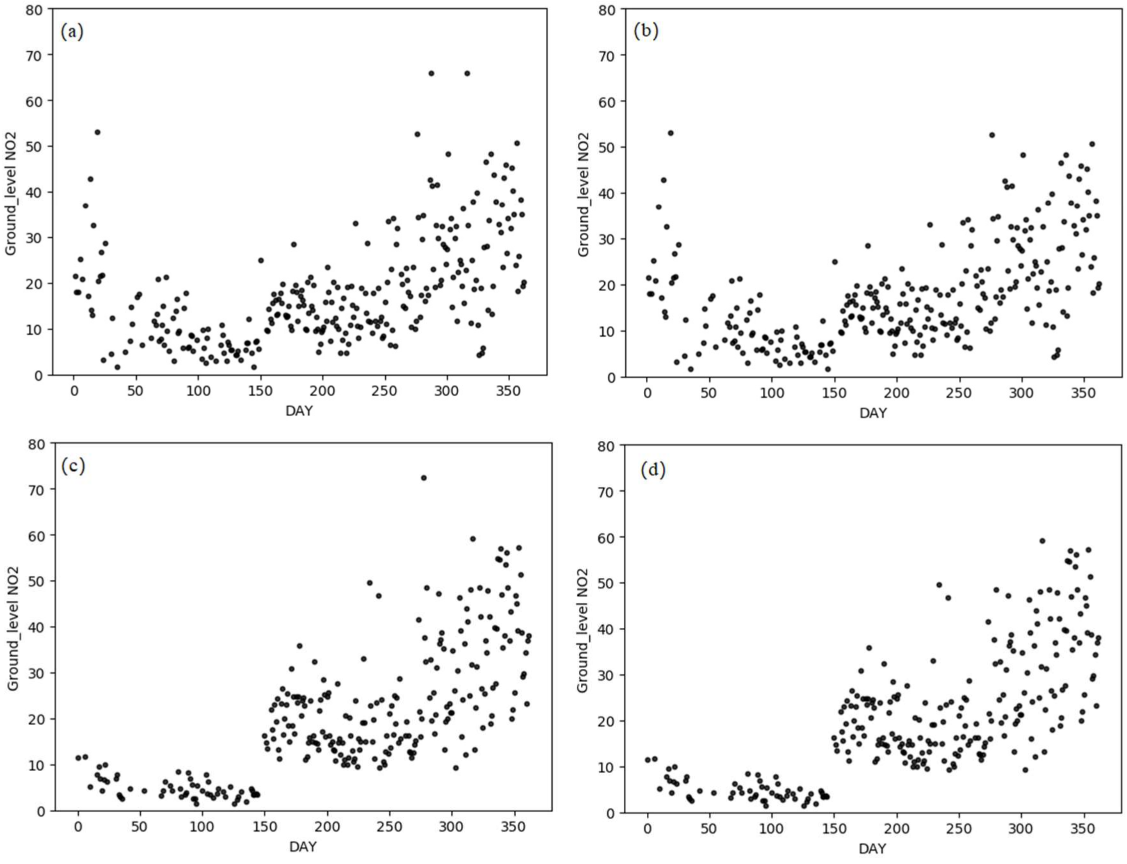


Figure 3. *Cont.*

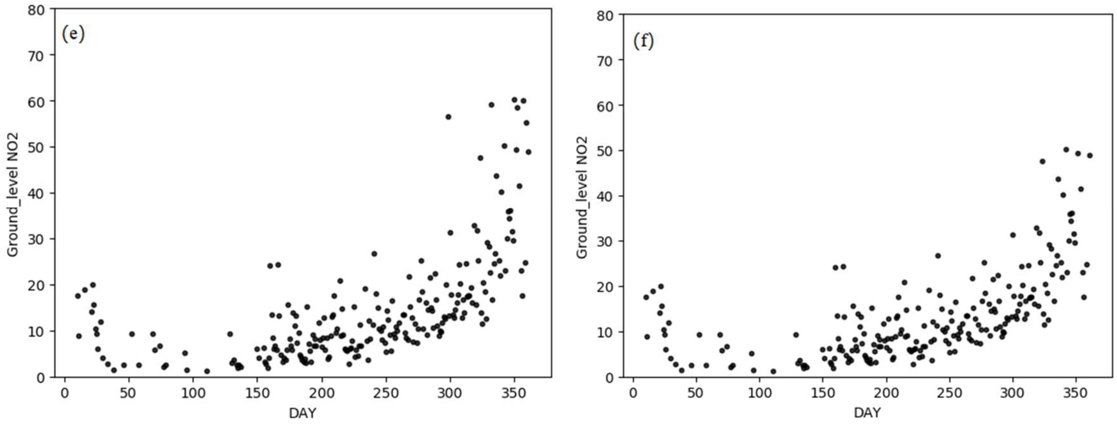


Figure 3. Comparison before (left-hand panels) and after (right-hand panels) quality control of ground sites. (**a**) before and (**b**) after quality control of the first site; (**c**) before and (**d**) after quality control of the second site; (**e**) before and (**f**) after quality control of the third site.

2.2.2. TropOMI Data Quality Control

As mentioned, TropOMI provides both a near-real-time and an offline tropospheric NO_2 product. The offline product can provide long time series of data that can be used to model training samples. However, the data are delayed by 10 days, meaning the established model cannot estimate the near-surface NO_2 concentration on the day of satellite observation. Conversely, the near-real-time data can be obtained 4 h after the detector scans a particular region on that day. However, the disadvantage is that this product cannot provide long-term data for modeling and generally only the data of the current month can be obtained from the website. Taking into account the relative strengths and limitations of the two products, it was decided that this study should use swath data (offline) for modeling. Then, the near-real-time data could be input into the established model to estimate the near-surface NO_2 concentration on that day.

The date of 16 August 2019 was selected randomly in this study to compare the two kinds of data in China, and data with quality assurance (qa)_values less than 0.75 for both kinds of data at the same time were excluded. The results are shown in Figure 4, from which we can see that the regions eliminated by the two kinds of data are basically consistent. It can further be seen from Table 4 that the maximum and average values of the two data types in China are basically the same, and the median difference is also very small, which is 0.156 mol/cm^2. In general, there is little difference between the spatial distribution and concentration distribution of the two kinds of data in China. The near real-time tropospheric NO_2 concentration data can be input into the model established by using the whole orbit, which makes the model highly effective.

Table 4. Statistic comparison of TropOMI's two types of tropospheric NO_2 column products.

	Maximum (mol/cm^2)	**Average (mol/cm^2)**	**Median (mol/cm^2)**
Near-real-time	40.529	1.168	1.031
Offline	39.715	1.012	0.875

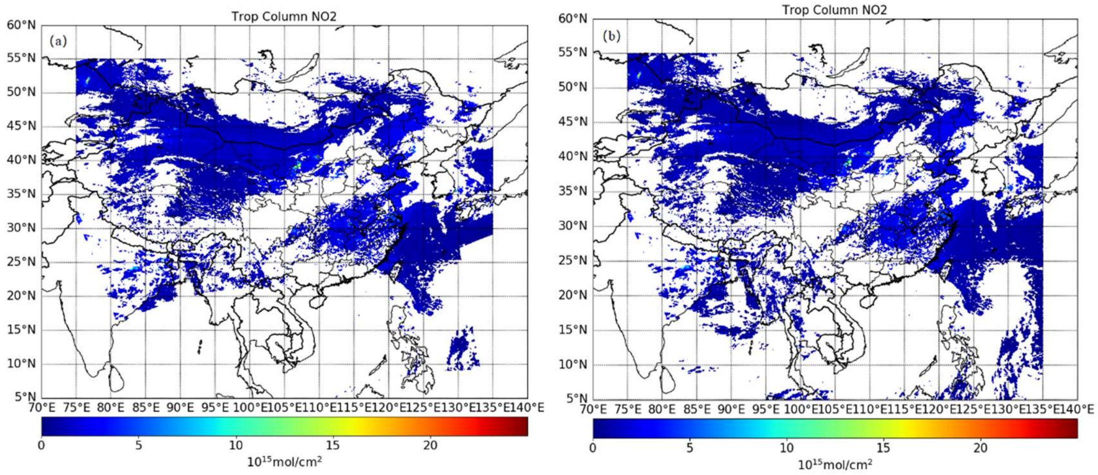


Figure 4. TropOMI tropospheric NO_2 column distribution in China on 16 August 2019: (**a**) near-real-time data; (**b**) offline data.

2.3. Data Set Generation

After the screening and quality control of the ground monitoring sites, a total of 963 sites were selected as target sites for modeling in China in 2019. The TropOMI tropospheric NO_2 products and meteorological forecast data provided by NCEP/GFS were selected according to these 963 ground sites in China.

Satellite data of tropospheric NO_2 with qa values > 0.75 were selected according to the TropOMI user manual. The meteorological data from the NCEP/GFS dataset predicted for 6 h at 00:00 (coordinated universal time) every day were selected and then changed to Beijing time to obtain the meteorological data at 14:00 each day. The timing of the satellite's scanning transit through China is 12:00–15:00 every day and the NCEP/GFS data are within this time period. The NCEP/GFS data used include the temperature at 2 m height, relative humidity, U and V wind components at 2 m height, and boundary layer height. Since the distribution of vegetation also has an impact on the NO_2 concentration, Level 1 data observed by MODIS were used to calculate the NDVI. The MODIS sensor passes over China at about 13:00 each day. According to the inverse distance weighting interpolation method, the NCEP/GFS forecast data and the NDVI data calculated by MODIS were matched for modeling near the 963 ground sites and within a 0.1° longitude and latitude range. When establishing the stochastic RF machine learning model, in addition to the above meteorological factors, the days of the modeling date in a year and the longitude and latitude information of the 963 monitoring sites after matching were also added.

Next, the ground monitoring site data on the 10th and 20th of each month in 2019 were selected, giving a total of 7959 effective values. These data were then used as matched sample data, all factors affecting their change were used as explanatory variables, and the daily NO_2 concentration of the ground site was used as the dependent variable to build an RF regression model. The model divides the training set and test set according to a ratio of 3:1. The divided training set has 5989 data values and the test set has 1990 data values.

3. Methods

3.1. Random Forest (RF) Algorithm Implementation

The RF method originated from the random decision forest method proposed by Breiman [46] in 2001, which is based on the earlier decision tree model and can rapidly classify data. However, due to it being unable to achieve the desired effect in follow-up machine learning research, Breiman combined the method with his own earlier idea of "bagging". He then carried out random repeated sampling in a range to ultimately obtain the newly proposed RF machine learning algorithm [47]. As an ensemble machine learning model, this algorithm needs to learn the subset model of the trained weak classifiers for many iterations, then combine these weak classifiers and finally vote on each subset or take their average value to obtain the final result of the model. RF can be divided into two types—a classified response variable (classifier) and a continuous response variable (regressor) [48]. RF is based on bootstrap sampling to train the samples to be learned. According to different classifiers, the best subset in each subsection is determined to divide the nodes in the decision tree. The principle of bootstrap sampling is based on random and repeated sampling, so the probability of each sample being sampled is the same. Even if the sample is not sampled in the last sampling, the probability it will still be sampled in the next sampling is the same. According to the calculation, in each extraction, the data with a probability of about 36.8%($\frac{1}{e}$) will not be extracted as samples, and this part of the data is also called "out of bag" data [48]. In addition, the algorithm can also give the importance of each explanatory variable. The higher the importance value, the more significant the explanatory variable is in estimating the target variable [49].

Due to the high accuracy and generalization of the model results of the overall learning of the RF learning model, it breaks through the single decision-making of the model determined by only one decision tree. Thus, the training method of one result is determined by multiple decision trees, which also makes the model less prone to over-fitting. This allows the prediction results to not only improve the accuracy but also to perform robustly, especially when there is high dimensionality in the data. Therefore, the machine learning model can better train the learning and predict the results after input [50]. In short, the prediction accuracy of the RF model is high [51], and it can deal with outliers and noise well. This explains the model's use in various scientific fields [52,53]. The RF algorithm computational package used in this study is provided by Scikit-Learn in Python, which contains the RF function [54–56].

In this study, a daily model for estimating the near-surface NO_2 concentration in China based on the RF algorithm was established and analyzed. The main steps are as follows: (1) establishing training and test data sets (Section 2: Data sets are created while selecting data and performing quality control.); (2) feature selection (Section 4.1); (3) using a machine learning library to build the platform of RF (Section 4.2); (4) model verification (Section 5).

3.2. Evaluation Procedure

Considering that the usual experimental evaluation methods have certain applicability and contingency, the ten-fold cross-validation method is used to test the accuracy of the model. That is, the dataset is divided into 10 parts, and then nine of them are taken in turn as the training data and the remaining one as the test data. The corresponding hit rate (or error rate) is obtained in each test, and the average value of the hit rate (or error rate) of the 10 results is used as an estimation of the accuracy of the algorithm.

The statistical metrics used in this study to measure the prediction performance of the developed model included: the coefficient of determination (R^2), RMSE, mean square error (MSE), and mean absolute error (MAE). The R^2 represents the relationship between random variables and multiple random variables, as follows:

$$R^2 = \frac{\sum(\hat{y} - \overline{y})^2}{\sum(y - \overline{y})^2} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2} \tag{2}$$

in which $\sum(y - \hat{y})^2$ is the sum of the squares of residuals, i.e., the sum of the squares of the deviation between the data points and their corresponding positions on the regression line, and $\sum(y - \overline{y})^2$ is the total sum of squares, i.e., the sum of the squares of the deviation between the data point and its average value. When R^2 is closer to 1, it indicates that the goodness-of-fit of the model is higher.

RMSE is a measure of the difference between the predicted value of the model and the real value. It is very sensitive to extremely large or small errors in a set of data, so it can reflect well the accuracy of the real value, as shown in Equation (3):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i' - y_i\right)^2} \tag{3}$$

On the other hand, MSE reflects the difference between the predicted value and the real value, which can be used to evaluate the degree of change in the model data. The smaller the value of the MSE, the higher the accuracy of the prediction model for the experimental data, as shown in Equation (4):

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i' - y_i\right)^2 \tag{4}$$

In addition, MAE is the average of the absolute value of the error between the predicted value and the real value of the model. This index can effectively avoid the problem of mutual offset between errors and accurately reflect the error of the prediction results, as shown in Equation (5):

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_i' - y_i\right| \tag{5}$$

In the above Equations (3)–(5), y_i' is the observed data, and y_i is the predicted data.

4. Establishment of Near-Surface NO$_2$ Concentration Estimation Model

4.1. Feature Selection

The RF model has the independence of multiple modeling and the randomness of variable selection. Moreover, it can compare the accuracy change degree of specific independent variables when they participate in modeling and when they do not participate in modeling. Therefore, it can reflect the importance score of each independent variable in regression. For the near-surface NO$_2$ concentration estimation, the TropOMI tropospheric NO$_2$ products and meteorological data (shown in Table 2) were herein selected as input features. After training the model with the training set, the importance score of each parameter in the model is output.

Figure 5 shows the importance ranking of adding explanatory variables. It can be seen that the tropospheric NO$_2$ concentration product retrieved by satellite is of key importance in the daily model for estimating the near-surface NO$_2$ concentration, being higher than 40%. This is followed by the influence of temperature at 2 m height and the days in a year, with the two basically having the same influence on the model. The *U* and *V* wind components and relative humidity contribute relatively less to the model.

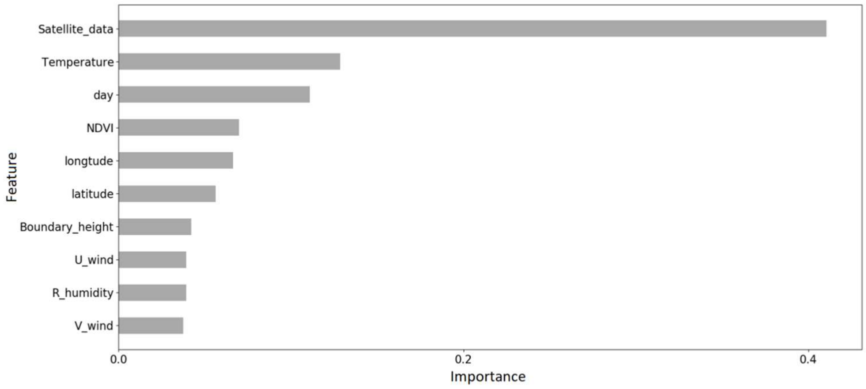
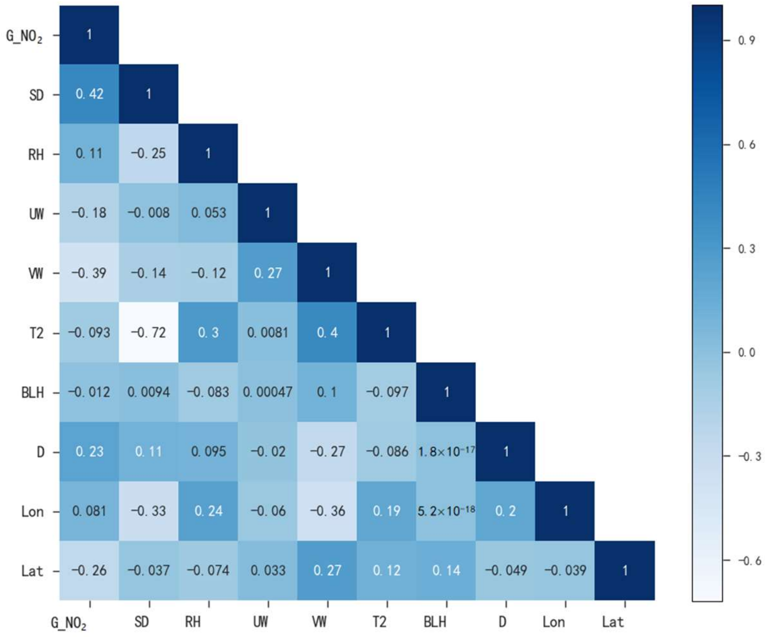


Figure 5. Ranking diagram of the importance of explanatory variables to the model.

The correlation between explanatory variables and dependent variable, i.e., the near-surface NO_2 concentration at the ground site, was calculated and the results are shown in Figure 6. It can be seen that the correlation between the ground NO_2 concentration and the satellite inversion product is the best, at 0.42, and the satellite product is also of primary importance in the model. The importance of the day factor is good and the correlation with the ground NO_2 concentration is 0.23. The V wind component has a good negative correlation with the ground NO_2 concentration, at -0.39. It can also be seen from Figure 6 that the correlation between explanatory variables is not high, and some variables have almost no relationship with each other. These results suggest that these explanatory variables meet the conditions of mutual independence and can be selected for model regression analysis.

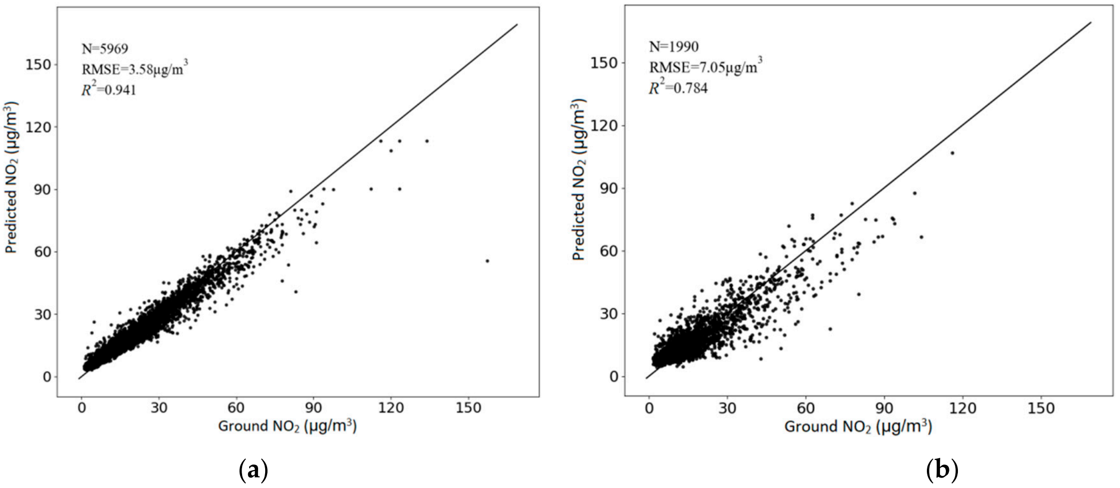


G_ NO_2 : the NO_2 concentration of the ground monitoring sites SD: the satellite data
RH: the relative humidity UW: the U component of the wind
VW: the V component of the wind T2: the temperature at 2 m height
BLH: the boundary layer height D: the number of days in a year
Lon: the longitude of the ground site Lat: the latitude of the ground site.

Figure 6. Distribution of correlation coefficients (CCs) between explanatory variables and dependent variables.

4.2. Construction of the RF Platform

The model verification index is used to judge the estimated results of the established near-surface NO_2 concentration model. Figure 7a shows the distribution of the correlation between predicted near-surface NO_2 concentration results by the RF model and measurements from ground sites in the model training set. Figure 7b is similar, but in the test set. Table 5 shows the model evaluation index values, such as R^2, RMSE, MSE, and so on, in the two data sets (training and test sets). From the figure, it can be seen that the results of the NO_2 daily concentration model training set are good, but there are still several large concentration points that have not yet been well studied, with an R^2 of 0.941 and RMSE of 3.58 μg/m^3. The R^2 of the NO_2 daily concentration model test set is 0.784 and the RMSE is 7.05 μg/m^3. Compared with the daily models of estimated near-surface NO_2 established by other researchers, such as Zhan et al. [57] (R^2 = 0.62; RMSE = 13.3 μg/m^3) and Chen et al. [32] (R^2 = 0.85; RMSE = 7.87 μg/m^3), it is clear that the RMSE result in this study is a significant improvement, but the R^2 still needs to be improved.



(**a**) (**b**)

Figure 7. The correlation between predicted near-surface NO_2 concentration results by RF model and measurements from ground sites: (**a**) training set and (**b**) test set.

Table 5. Evaluation results of the RF model in the training set and test set.

	Number of Samples	R^2	RMSE (μg/m^3)	MSE (μg/m^3)	MAE (μg/m^3)	Interpretation Degree
Training set	5969	0.94	3.59	12.90	2.21	0.94
Test set	1990	0.78	7.04	49.54	5.00	0.78

5. Results and Discussion

Based on the RF machine learning technique, a daily model for estimating the near-surface NO_2 concentration was established by using the NO_2 concentration data of ground sites and meteorological data in 2019. The test set verified that the proposed model produces satisfactory predictions. Therefore, this model was further used to analyze the spatiotemporal distribution of the near-surface NO_2 concentration in China during the period of COVID-19 in early 2020. Results verification was performed with monitoring data from ground sites in the same period.

The region of China from 1–5 February 2019 and 1–5 February 2020 was separately selected for estimation of near-surface NO_2 concentrations, and the results are shown in Figure 8. As can be seen by comparing the two sets of results, due to the impact of COVID-19, when the government called upon citizens to stay at home and reduce their travel, the near-surface NO_2 concentration caused by human travel was alleviated. Compared to the same period in 2019, the area polluted by NO_2 was significantly reduced. From 1–5 February 2019, there are obvious areas with serious pollution—in Beijing, Tianjin and Hebei, northeast and southwest China—and the maximum concentration is close to 40 μg/m^3. However, in the same period in 2020, the near-surface NO_2 concentration in these areas is significantly lower; it basically did not exceed 30 μg/m^3. Notably, though, the level of NO_2 pollution in the Xinjiang area did not change much compared with the previous years. From the estimated distribution of NO_2 concentration near the ground in China, it can be seen that the concentration of NO_2 pollution in China decreased significantly in February 2020. This is reflective of China's timely and effective deployment of prevention and control measures following the sudden outbreak of COVID-19. From 1–5 February 2020, it can be seen that the near-surface NO_2 concentration in polluted areas of China decreased day by day, to the point where the NO_2 concentration was basically around 20 μg/m^3 on 5 February 2020, which is a relatively low level. The estimated results of the model show a significantly different situation in the east of Northeast China and the southwest of Xinjiang and Tibet. This may be because the sites in these areas are not screened-out during modeling, resulting in poor machine learning and rendering these areas consistent with the surrounding pollution. In the western region, where there are few ground monitoring sites for geographical reasons, the model can be used to estimate the near-surface NO_2 concentration, thus compensating for the lack of observations and thereby providing data support for pollution research in this region.

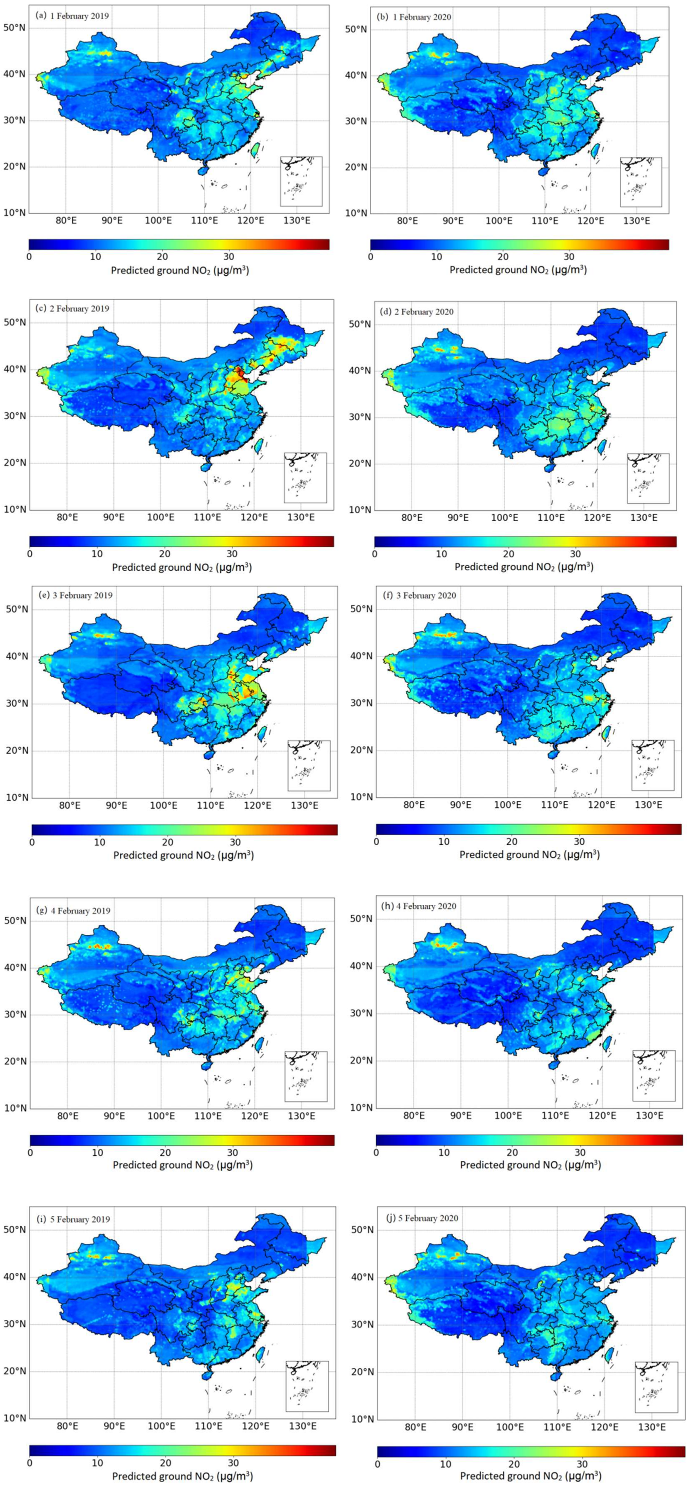


Figure 8. Estimated near-surface NO_2 concentrations by the RF model in China from 1 to 5 February 2019 and 2020: (**a**) 1 February 2019; (**b**) 1 February 2020; (**c**) 2 February 2019; (**d**) 2 February 2020; (**e**) 3 February 2019; (**f**) 3 February 2020; (**g**) 4 February 2019; (**h**) 4 February 2020; (**i**) 5 February 2019; (**j**) 5 February 2020.

Figure 9 illustrates the estimated average value of the near-surface NO_2 concentration in seven typical cities in China from 1–5 February 2019 and 2020. It can be seen that the near-surface NO_2 concentration in these seven cities shows different changes. Among them, the NO_2 concentration in Beijing, Wuhan, Nanjing, Zhengzhou and Chengdu in 2019 is higher than that in 2020, especially in Chengdu, where the change is the most obvious, with a decrease of 14.56 μg/m^3. On the other hand, it is noted that the concentration in Shanghai and Urumqi in 2020 is slightly higher than that in 2019. The near-surface NO_2 concentration in Urumqi in 2020 is 5.11 μg/m^3 higher than that in 2019.

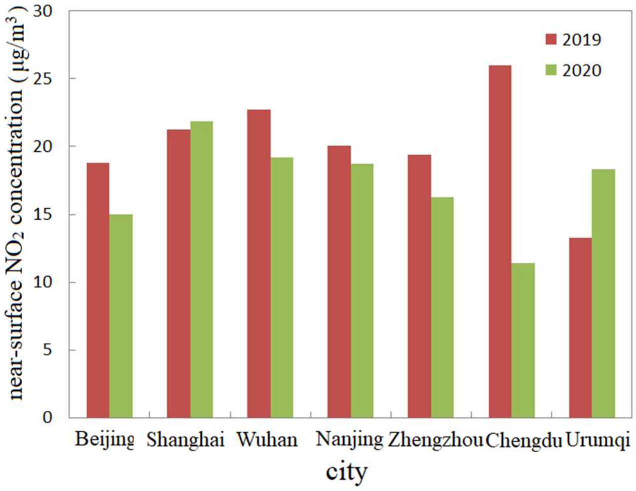
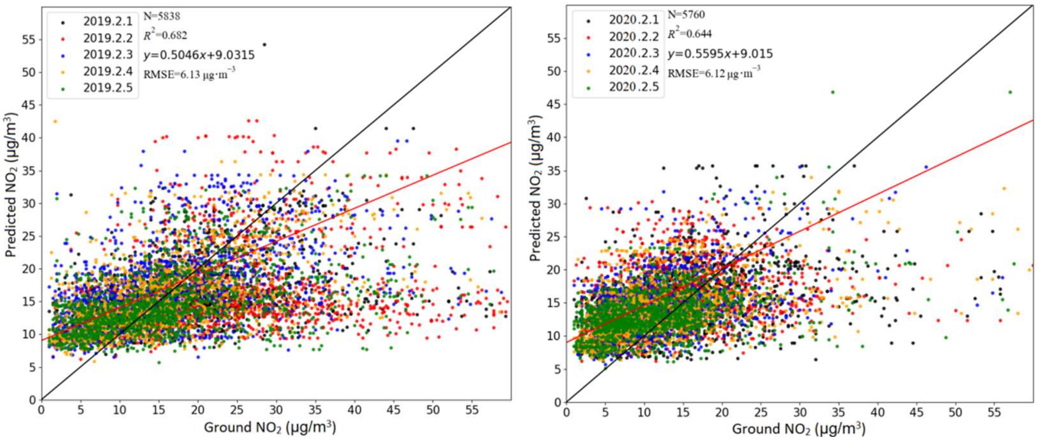


Figure 9. Near-surface NO_2 concentration comparison in seven typical cities in China between 1 to 5 February 2019 and 2020.

The predicted near-surface NO_2 concentrations were matched with the monitoring results of the ground sites in China. "Matched" here means the matching method for geographical location involved matching the satellite data at each ground site by using the inverse distance weighting interpolation method within a 0.1° longitude and latitude range of the ground site. As for time matching, since TropOMI passes through China from 12:00 to 15:00 (Beijing time) every day, the hourly data of the ground sites in this period were selected to calculate the average concentration as the NO_2 concentration of a particular site on that day to match with the satellite products. The predicted values of the NO_2 concentration model were selected according to the inverse distance weighting interpolation method and the correlation between them was calculated. As an example, Figure 10 shows the correlation between the NO_2 concentration in China for five days in 2019 and 2020. In particular, Figure 10a shows the correlation between them in 2019 and Figure 10b shows the correlation between them in 2020. From 1–5 February 2019, there are 5938 data samples in total, with an R^2 of 0.682 and MSE of 6.13 μg/m^3. From 1–5 February 2020, there are 5760 data samples in total, with an R^2 of 0.644 and MSE of 6.12 μg/m^3. In terms of correlation, the results estimated by the model in 2019 and 2020 are to a certain extent correlated with the concentrations monitored on the ground, The 2019 validation results are better than 2020, but the results are not much different. This may be due to the better generalization ability of the model, but there are still deviations between the estimated values of the model and the actual monitored values at some sites; multi-year data can be used for modeling to increase the model's learning of the differences between different years.



(**a**) (**b**)

Figure 10. The correlation between predicted near-surface NO$_2$ concentration results by RF model and measurements from ground sites from 1–5 February 2019 and 2020: (**a**) 1–5 February 2019; (**b**) 1–5 February 2020.

6. Conclusions

In this study, TropOMI-observed tropospheric NO$_2$ column data was first used, compiled with other auxiliary data with the aim of establishing a daily near-surface NO$_2$ concentration estimation model by RF in China. The key study findings are summarized below as follows:

1. In the daily model established to estimate the near-surface NO$_2$ concentration in China based on the RF method, the tropospheric NO$_2$ columns observed by TropOMI have a substantial influence on the importance of the model and the highest correlation with the near-surface NO$_2$ concentration. It can be seen that adding satellite data into the model is helpful in estimating near-surface NO$_2$ concentrations. Furthermore, temperature, date and NDVI also contribute to the NO$_2$ concentration estimation near the surface.
2. The established model was verified based on the ten-fold cross-validation method. The R^2 and RMSE of the model for estimating the daily NO$_2$ concentration in China by the RF model are 0.78 and 7.04 μg/m^3, respectively. Compared with previous studies on modeling the daily NO$_2$ concentration, there are still some gaps in the R^2 but the RMSE is an improvement.
3. The established model was used to estimate the near-surface NO$_2$ concentration in China from 1–5 February 2020. Compared to the same period in 2019, it was found that the near-surface NO$_2$ concentration in most parts of China decreased significantly in 2020, especially in the Beijing–Tianjin–Hebei region and Fenwei plain. This suggests that the strong measures taken by the Chinese government to control the COVID-19 pandemic have been well reflected by the forecast model, reflecting the practicability of the model forecast. From an analysis of seven typical cities, apart from Shanghai and Urumqi where the near-surface NO$_2$ concentration in 2020 was slightly higher than that in 2019, other cities showed an obvious downward trend. By comparing and verifying the model-estimated results with the ground monitoring results, it was found that the results in 2019 and 2020 were basically the same. The R^2 values reported were 0.682 and 0.644, and the RMSEs were 6.13 and 6.12 μg/m^3, respectively. It further verifies that the model has good practicability in China.

All in all, the present study demonstrated the use of the established daily model using the newly proposed RF machine learning method for estimating the near-surface NO$_2$ concentration in China. Furthermore, it was shown that the distribution of near-surface NO$_2$ could be obtained on the same day. Our proposed model offers satisfactory timeliness and improved RMSE from 7.87 to 7.05 μg/m^3 compared with previous studies.

It provides an appropriate solution to be used in cases where there is lack of observations in some geographically challenging areas of China. Future directions to continue this study could include using multiple satellite products to establish a long-term near-surface NO_2 daily concentration model for China, and also adding an emissions inventory to further improve the model. In addition, it is also necessary to compare the RF outputs with other machine learning (ML) methods such as SVM and deep learning. This study can provide data support for the Chinese environmental protection department to formulate relevant policies, as well as provide some help for the evaluation of the simulation results of atmospheric chemistry models.

Author Contributions: Conceptualization, Y.W.; methodology, Y.W. and M.L.; software, M.L. and B.L.; validation, M.L. and B.L.; formal analysis, Y.W., Y.B. and G.P.P.; investigation, M.L. and B.L.; resources, Y.B.; data curation, M.L. and B.L.; writing—original draft preparation, Y.W. and G.P.P.; writing—review and editing, G.P.P.; visualization, M.L. and B.L.; supervision, Y.B.; project administration, Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Program of China (2017YFC1501704) and National Natural Science Foundation of China (41975046) and NUIST Students' Platform for Innovation and Entrepreneurship Training Program (202210300113Y, Reading2022003).

Data Availability Statement: The data supporting this research can be found at the hyperlinks: https://s5phub.copernicus.eu/dhus/#/home (accessed on 12 March 2020); https://rda.ucar.edu (accessed on 16 March 2020); http://modis.gsfc.nasa.gov (accessed on 20 March 2020); http://www.cnemc.cn (accessed on 25 March 2020).

Acknowledgments: We thank the anonymous reviewers for the comments that helped improve our manuscript.

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

References

1. Logan, J.A. Nitrogen Oxides in the troposphere: Global and Regional Budgets. *J. Geophys. Res.* **1983**, *88*, 10785–10807. [CrossRef]
2. Solomon, S.; Portmann, R.W.; Sanders, R.W.; Daniel, J.S.; Madsen, W.; Bartram, B.; Dutton, E.G. On The Role of Nitrogen Dioxide in the Absorption of Solar Radiation. *J. Geophys. Res.* **1999**, *1041*, 12047–12058. [CrossRef]
3. Stavrakou, T.; MÜLler, J.F.; Boersma, K.F.; Van der, A.R.J.; Kurokawa, J.; Ohara, T.; Zhang, Q. Key Chemical NO_x Sink Uncertainties and how They Influence Top-Down Emissions of Nitrogen Oxides. *Atmos. Chem. Phys.* **2013**, *13*, 7871–7929. [CrossRef]
4. Crutzen, P.J. The Role of NO and NO_2 in the Chemistry of the Troposphere and Stratosphere. *Annu. Rev. Earth Planet. Sci.* **1979**, *7*, 443–472. [CrossRef]
5. Volz, A.; Kley, D. Evaluation of the Montsouris Series of Ozone Measurements Made in the Nineteenth Century. *Nature* **1988**, *332*, 240–242. [CrossRef]
6. Barbara, J.; Finlayson-Pitts, J.N.; Pitts, J. CHAPTER 4—Photochemistry of Important Atmospheric Species. In *Chemistry of the Upper and Lower Atmosphere*; Elsevier: Amsterdam, The Netherlands, 2000; pp. 86–129.
7. Duncan, B.N.; Lamsal, L.N.; Thompson, A.M.; Yoshida, Y.; Lu, Z.; Streets, D.G.; Hurwitz, M.M.; Pickering, K.E. A Space-Based, High-Resolution View of Notable Changes in Urban NO_x Pollution around the World (2005–2014). *J. Geophys. Res. Atmos.* **2016**, *121*, 976–996. [CrossRef]
8. Samoli, E. Short-Term Effects of Nitrogen Dioxide on Mortality: An Analysis within the APHEA Project. *Eur. Respir. J.* **2006**, *27*, 1129–1138. [CrossRef]
9. Latza, U.; Gerdes, S.; Baur, X. Effects of Nitrogen Dioxide on Human Health: Systematic Review of Experimental and Epidemiological Studies Conducted between 2002 and 2006. *Int. J. Hyg. Environ. Health* **2009**, *212*, 271–287. [CrossRef] [PubMed]
10. Chen, R.; Samoli, E.; Wong, C.M.; Huang, W.; Wang, Z.; Chen, B.; Kan, H. CAPES Collaborative Group. Associations between Short-Term Exposure to Nitrogen Dioxide and Mortality in 17 Chinese Cities: The China Air Pollution and Health Effects Study (CAPES). *Environ. Int.* **2012**, *45*, 32–38. [CrossRef]
11. Zhang, R.; Li, Q.; Zhang, R. Meteorological Conditions for the Persistent Severe Fog and Haze Event over Eastern China in January 2013. *Sci. China Earth Sci.* **2014**, *44*, 26–35.
12. Leue, C.; Wenig, M.; Wagner, T.; Klimm, O.; Platt, U.; Jhne, B. Quantitative Analysis of NO_x Emissions from Global Ozone Monitoring Experiment Satellite Image Sequences. *J. Geophys. Res. Atmos.* **2001**, *106*, 5493–5505. [CrossRef]
13. Shi, C.E.; Zhang, B.N. Tropospheric NO_2 Columns over Northeastern North America: Comparison of CMAQ Model Simulations with GOME Satellite Measurements. *Adv. Atmos. Sci.* **2008**, *1*, 59–71. [CrossRef]

14. Lamsal, L.N.; Duncan, B.N.; Yoshida, Y.; Krotkov, N.A.; Pickering, K.E.; Streets, D.G.; Lu, Z.U.S. NO_2 Trends (2005–2013): EPA Air Quality System (AQS) Data versus Improved Observations from the Ozone Monitoring Instrument (OMI). *Atmos. Environ.* **2015**, *110*, 130–143. [CrossRef]
15. Lamsal, L.N.; Martin, R.V.; Van Donkelaar, A.D.; Steinbacher, M.; Celarier, E.A.; Bucsela, E.; Dunlea, E.J.; Pinto, J.P. Ground-Level Nitrogen Dioxide Concentrations Inferred from the Satellite-Borne Ozone Monitoring Instrument. *J. Geophys. Res. Atmos.* **2008**, *113*, 1–15. [CrossRef]
16. Levelt, P.F.; Hilsenrath, E.; Leppelmeier, G.W.; Van Den Oord, G.H.J.; Bhartia, P.K.; Tamminen, J.; De Haan, J.F.; Veefkind, J.P. Science Objectives of the Ozone Monitoring Instrument. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1199–1208. [CrossRef]
17. Levelt, P.F.; Van Den Oord, G.H.J.; Dobber, M.R.; Malkki, A.; Visser, H.; De Vries, J.; Stammes, P.; Lundell, J.O.V.; Saari, H. The Ozone Monitoring Instrument. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1093–1101. [CrossRef]
18. Martin, R.V.; Parrish, D.D.; Ryerson, T.B.; Nicks, D.K., Jr.; Chance, K.; Kurosu, T.P.; Jacob, D.J.; Sturges, E.D.; Fried, A.; Wert, B.P. Evaluation of GOME Satellite Measurements of Tropospheric NO_2 and HCHO Using Regional Data from Aircraft Campaigns in the Southeastern United States. *J. Geophys. Res. Atmos.* **2004**, *109*, D24307. [CrossRef]
19. Bucsela, E.J.; Perring, A.E.; Cohen, R.C.; Boersma, K.F.; Celarier, E.A.; Gleason, J.F.; Wenig, M.O.; Bertram, T.H.; Wooldridge, P.J.; Dirksen, R. Comparison of Tropospheric NO_2 from in Situ Aircraft Measurements with Near-Real-Time and Standard Product Data from OMI. *J. Geophys. Res. Atmos.* **2008**, *113*, D16S31. [CrossRef]
20. Burrows, J.P.; Weber, M.; Buchwitz, M.; Rozanov, V.; Ladstätter-Weißenmayer, A.; Richter, A.; Debeek, R.; Hoogen, R.; Bramstedt, K.; Eichmann, K.; et al. The Global Ozone Monitoring Experiment (GOME): Mission Concept and First Scientific Results. *J. Atmos. Sci.* **1999**, *56*, 151–175. [CrossRef]
21. Bovensmann, H.; Burrows, J.P.; Buchwitz, M.; Frerick, J.; Noël, S.; Rozanov, V.V.; Chance, K.V.; Goede, A.P.H. SCIAMACHY: Mission Objectives and Measurement Modes. *J. Atmos. Sci.* **1999**, *56*, 127–150. [CrossRef]
22. Zhao, X.; Fioletov, V.; Alwarda, R.; Su, Y.; Griffin, D.; Weaver, D.; Strong, K.; Cede, A.; Hanisco, T.; Tiefengraber, M.; et al. Tropospheric and Surface Nitrogen Dioxide Changes in the Greater Toronto Area during the First Two Years of the COVID-19 Pandemic. *Remote Sens.* **2022**, *14*, 1625. [CrossRef]
23. Zhang, Q.; Streets, D.G.; He, K.; Wang, Y.; Richter, A.; Burrows, J.P.; Uno, I.; Jang, C.J.; Chen, D.; Yao, Z.; et al. NO_x Emission Trends for China, 1995–2004: The View from the Ground and the View from Space. *J. Geophys. Res. Atmos.* **2007**, *112*, D22306. [CrossRef]
24. Petritoli, A.; Bonasoni, P.; Giovanelli, G.; Ravegnani, F.; Kostadinov, I.; Bortoli, D.; Weiss, A.; Schaub, D.; Richter, A.; Fortezza, F. First Comparison between Ground-Based and Satellite-Borne Measurements of Tropospheric Nitrogen Dioxide in the Po Basin. *J. Geophys. Res. Atmos.* **2004**, *109*, D15307. [CrossRef]
25. OrdÓNez, C.; Richter, A.; Steinbacher, M.; Zellweger, C.; Prévot, A.S.H.; Ordónez, C.; Nü, B.; Burrows, J.P. Comparison of 7 Years of Satellite-Borne and Ground-Based Tropospheric NO_2 Measurements around Milan, Italy. *J. Geophys. Res. Atmos.* **2006**, *111*, D05310. [CrossRef]
26. Qin, K.; Rao, L.; Xu, J.; Bai, Y.; Zou, J.; Hao, N.; Li, S.; Yu, C. Estimating Ground Level NO_2 Concentrations over Central-Eastern China Using A Satellite-Based Geographically and Temporally Weighted Regression Model. *Remote Sens.* **2017**, *9*, 950. [CrossRef]
27. Liu, F.; Beirle, S.; Zhang, Q.; Dörner, S.; He, K.; Wagner, T. NO_x Lifetimes and Emissions of Cities and Power Plants in Polluted Background Estimated by Satellite Observations. *Atmos. Chem. Phys.* **2015**, *15*, 24179–24215.
28. Bechle, M.J.; Millet, D.B.; Marshall, J.D. Remote Sensing of Exposure to NO_2: Satellite versus Ground-Based Measurement in a Large Urban Area. *Atmos. Environ.* **2013**, *69*, 345–353. [CrossRef]
29. Kim, H.C.; Lee, S.-M.; Chai, T.; Ngan, F.; Pan, L.; Lee, P. A Conservative Downscaling of Satellite-Detected Chemical Compositions: NO_2 Column Densities of OMI, GOME-2, and CMAQ. *Remote Sens.* **2018**, *10*, 1001. [CrossRef]
30. Byun, D.W.; Ching, J. *Science Algorithms of the EPA Models-3 Community Multi-Scale Air Quality (CMAQ) Modeling System*; NERL: Research Triangle Park, NC, USA, 1999; p. 425.
31. Ghahremanloo, M.; Lops, Y.; Choi, Y.; Yeganeh, B. Deep Learning Estimation of Daily Ground-Level NO_2 Concentrations From Remote Sensing Data. *J. Geophys. Res. Atmos.* **2021**, *126*, e2021JD034925. [CrossRef]
32. Chen, Z.Y.; Zhang, R.; Zhang, T.H.; Ou, C.Q.; Guo, Y. A Kriging-Calibrated Machine Learning Method for Estimating Daily Ground-Level NO_2 in Mainland China. *Sci. Total Environ.* **2019**, *690*, 556–564. [CrossRef] [PubMed]
33. Lee, H.J.; Koutrakis, P. Daily ambient NO_2 concentration predictions using satellite ozone monitoring instrument NO_2 data and land use regression. *Environ. Sci. Technol.* **2014**, *48*, 2305–2311.
34. Maasakkers, J.D.; Boersma, K.F.; Williams, J.E.; Van Geffen, J.; Veefkind, J.P. Vital Improvements to the Retrieval of Tropospheric NO_2 Columns from the Ozone Monitoring Instrument. In Proceedings of the EGU General Assembly Conference Abstracts, Vienna, Austria, 7–12 April 2013.
35. Wang, C.; Wang, T.; Wang, P.; Rakitin, V.S. Comparison and Validation of TROPOMI and OMI NO_2 Observations over China. *Atmosphere* **2020**, *11*, 636. [CrossRef]
36. Wang, Y.; Wang, J.; Zhou, M.; Henze, C.D.K.; Wang, W.G. Inverse Modeling of SO_2 and NO_x; Emissions over China Using Multisensor Satellite Data Part 2: Downscaling Techniques for Air Quality Analysis and Forecasts. *Atmos. Chem. Phys.* **2020**, *20*, 6651–6670. [CrossRef]

37. Theys, N.; Smedt, I.D.; Yu, H.; Danckaert, T.; Gent, J.V.; Hörmann, C.; Wagner, T.; Hede-lt, P.; Bauer, H.; Romahn, F. Sulfur Dioxide Retrievals from TROPOMI Onboard Sentinel-5 Precursor: Algorithm Theoretical Basis. *Atmos. Meas. Tech.* **2017**, *10*, 119–153. [CrossRef]
38. Borsdorff, T.; Andrasec, J.; Joost, A.D.B.; Hu, H.; Aben, I.; Landgraf, J. Detection of Carbon Monoxide Pollution from Cities and Wildfires on Regional and Urban Scales: The Benefit of CO Column Retrievals from SCIAMACHY 2.3MM Measurements Under Cloudy Conditions. *Atmos. Meas. Tech.* **2018**, *11*, 2553–2565. [CrossRef]
39. Van Geffen, J.H.G.M.; Boersma, K.F.; Van Roozendael, M.; Hendrick, F.; Mahieu, E.; Smed, I.D.; Sneep, M.; Veefkind, J.P. Improved Spectral Fitting of Nitrogen Dioxide from OMI in the 405–465 Nm Window. *Atmos. Meas. Tech.* **2015**, *8*, 1685–1699. [CrossRef]
40. Torres, O.; Decae, R.; Veefkind, J.P.; de Leeuw, G. OMI Aerosol Retrieval Algorithm. In *OMI Algorithm Theoretical Basis Docu-ment: Clouds, Aerosols, and Surface UV Irradiance, 3, Version 2, OMI-ATBD-03*; Stammes, P., Ed.; NASA Goddard Space Flight Center: Greenbelt, MD, USA, 2002; pp. 47–71.
41. Gu, J.; Chen, L.; Yu, C.; Li, S.; Tao, J.; Fan, M.; Xiong, X.; Wang, Z.; Shang, H.; Su, L. Ground-Level NO_2 Concentrations over China Inferred from the Satellite OMI and CMAQ Model Simulations. *Remote Sens.* **2017**, *9*, 519. [CrossRef]
42. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
43. Waskom, M.L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
44. Bisong, E. *Matplotlib and Seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Apress: Berkeley, CA, USA, 2019; pp. 151–165.
45. Pukelsheim, F. The Three Sigma Rule. *Am. Stat.* **1994**, *48*, 88–91.
46. Breiman, L. Random Forests, Machine Learning 45. *J. Clin. Microbiol.* **2001**, *2*, 199–228.
47. Wu, H.; Ying, W. Benchmarking Machine Learning Algorithms for Instantaneous Net Surface Shortwave Radiation Retrieval Usingremote Sensing Data. *Remote Sens.* **2019**, *11*, 2520. [CrossRef]
48. Liaw, A.; Wiener, M. Classification and Regression by Random-Forest. *R News* **2002**, *23*, 18–22.
49. Scornet, E. Random Forests and Kernel Methods. *IEEE Trans. Inf. Theory* **2016**, *62*, 1485–1500. [CrossRef]
50. Lee, S.; Kim, J. Prediction of Nanofiltration and Reverse-Osmosis-Membrane Rejection of Organic Compounds Using Random Forest Model. *J. Environ. Eng.* **2020**, *146*, 04020127.
51. Lee, Y.; Han, D.; Ahn, M.H.; Im, J.; Lee, S.J. Retrieval of Total Precipitable Water from Himawari-8 AHI Data: A Comparison of Random Forest, Extreme Gradient Boosting, and Deep Neural Network. *Remote Sens.* **2019**, *11*, 1741. [CrossRef]
52. Jenkins, J.M.; Kowalski, M.; Alvarenga, E.F.S. Predictive Modelling of Water Losses Using Random Forests on Weather Covariates. *Water Supply* **2018**, *18*, 2180–2187.
53. Ruan, C.; Dong, Y.; Huang, W.; Huang, L.; Ye, H.; Ma, H.; Guo, A.; Sun, R. Integrating Remote Sensing and Meteorological Data to Predict Wheat Stripe Rust. *Remote Sens.* **2022**, *14*, 1221. [CrossRef]
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Bressert, E. *SciPy and NumPy: An Overview for Developers*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2012.
56. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. In Proceedings of the ECML PKDD Workshop: Languages for Data Mining and Machine Learning, Bristol, UK, 23–27 September 2013; pp. 108–122.
57. Zhan, Y.; Luo, Y.; Deng, X.; Zhang, K.; Zhang, M.; Grieneisen, M.L.; Di, B. Satellite-Based Estimates of Daily NO_2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model. *Environ. Sci. Technol.* **2018**, *52*, 4180–4189. [PubMed]