remote sensing

MDPI

Technical Note

Satellite-Based Estimation of Near-Surface NO_2 Concentration in Cloudy and Rainy Areas

Fuliang Deng [1], Yijian Chen [1], Wenfeng Liu [1], Lanhui Li [1], Xiaojuan Chen [2], Pravash Tiwari [2] and Kai Qin [2,*]

[1] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; 2222031137@s.xmut.edu.cn (W.L.); lilh@xmut.edu.cn (L.L.)
[2] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China
* Correspondence: qinkai@cumt.edu.cn

Abstract: Satellite-based remote sensing enables the quantification of tropospheric NO_2 concentrations, offering insights into their environmental and health impacts. However, remote sensing measurements are often impeded by extensive cloud cover and precipitation. The scarcity of valid NO_2 observations in such meteorological conditions increases data gaps and thus hinders accurate characterization and variability of concentration across geographical regions. This study utilizes the Empirical Orthogonal Function interpolation in conjunction with the Extreme Gradient Boosting (XGBoost) algorithm and dense urban atmospheric observed station data to reconstruct continuous daily tropospheric NO_2 column concentration data in cloudy and rainy areas and thereby improve the accuracy of NO_2 concentration mapping in meteorologically obscured regions. Using Chengdu City as a case study, multiple datasets from satellite observations (TROPOspheric Monitoring Instrument, TROPOMI), near-surface NO_2 measurements, meteorology, and ancillary data are leveraged to train models. The results showed that the integration of reconstructed satellite observations with provincial and municipal control surface measurements enables the XGBoost model to achieve heightened predictive accuracy (R^2 = 0.87) and precision ($RMSE$ = 5.36 μg/m^3). Spatially, this approach effectively mitigates the problem of missing values in estimation results due to absent satellite data while simultaneously ensuring increased consistency with ground monitoring station data, yielding images with more continuous and refined details. These results underscore the potential for reconstructing satellite remote sensing information and combining it with dense ground observations to greatly improve NO_2 mapping in cloudy and rainy areas.

Keywords: TROPOMI; near-surface NO_2 concentration; municipal control surface measurements; machine learning; data reconstruction

Citation: Deng, F.; Chen, Y.; Liu, W.; Li, L.; Chen, X.; Tiwari, P.; Qin, K. Satellite-Based Estimation of Near-Surface NO_2 Concentration in Cloudy and Rainy Areas. *Remote Sens.* **2024**, *16*, 1785. https://doi.org/10.3390/rs16101785

Academic Editor: Carmine Serio

Received: 20 March 2024
Revised: 13 May 2024
Accepted: 13 May 2024
Published: 17 May 2024

1. Introduction

Nitrogen dioxide (NO_2), as a pivotal trace gas within the atmosphere [1], exerts a significant impact on the ecological environment and climate change [2–4]. As an important indicator of local air quality, near-surface NO_2 concentration exhibits spatial and temporal heterogeneity associated with proximity to emission sources that can aggregate into concentrated zones, posing health risks [5–7]. Currently, China has established over 1500 ground-level air quality monitoring stations to measure near-surface NO_2 concentrations, while uneven distributions of ground stations engender uncertainty in quantifying exposure risks and associated health burdens, especially in sparsely sampled outlying populations [8,9]. Hence, a more refined and accurate estimation of near-surface NO_2 concentration is of critical importance for accurately delineating urban air quality patterns to inform precise atmospheric environmental regulation.

The continual advancement of satellite remote sensing technology has enabled increased spatial observation of NO_2 concentration, compensating coverage gaps in the

surface monitoring network [10], and at a lower expense than that associated with aircraft surveys [11]. Chemical transport models that establish a relationship between surface concentrations and satellite-based observations have traditionally been used to determine ground-level NO_2 concentrations. Initially, Lamsal et al. [12] employed scaling factors from the global three-dimensional model (GEOS-CHEM) in conjunction with tropospheric NO_2 column concentrations from the Ozone Monitoring Instrument (OMI) to derive near-surface NO_2 concentrations over North America. Their findings indicated a significant correlation between these measurements. However, this method is constrained by its lower precision and the limited spatial resolution of the derived NO_2 concentration data. Furthermore, diverse methodologies, spanning traditional regression models to machine learning techniques, have been utilized in conjunction with satellite data for estimation purposes [10]. The Land Use Regression (LUR) model, in particular, has demonstrated notable success in estimating near-surface NO_2 concentrations across various countries, including the United States, Canada [13], Australia [14], and the United Kingdom [15], and even on a global scale [16].

Since the adoption and release of the new version of "Environmental Air Quality Standards" (GB 3095-2012) [17] in 2012 and the proposal of the "Blue Sky Defense" plan at the fifth session of the 12th National People's Congress of the People's Republic of China in 2017, satellite-based estimation of near-surface NO_2 concentrations has attracted increasing attention. Scholars have been leveraging regression models to estimate near-surface NO_2 concentrations using satellite data in combination with ground monitoring station data, and recent studies have demonstrated that the ground-level estimations of near-surface NO_2 concentrations achieved meaningful progress. Qin et al. [18] utilized four methods—Geographically and Temporally Weighted Regression (GTWR), Ordinary Least Squares (OLS), Geographically Weighted Regression (GWR), and Temporally Weighted Regression (TWR)—to estimate the near-surface NO_2 concentration in Eastern China based on OMI satellite NO_2 data and meteorological data, with GTWR demonstrating the highest estimation accuracy (R^2 = 0.60). With the evolving applications of artificial intelligence, machine learning methods have shown superior predictive capabilities in estimating near-surface NO_2 concentrations [19]. Araki et al. [20] estimated the concentration of near-surface NO_2 in Japan using a Land Use Random Forest model (LURF) (R^2 = 0.79), outperforming traditional LUR models. You et al. [21] conducted estimations of China's near-surface NO_2 concentration based on the random forest algorithm and multi-source geospatial data, achieving a monthly scale model (R^2 = 0.84), which surpassed the estimation using the LUR model. Chi et al. [22] employed the Extreme Gradient Boosting (XGBoost) machine learning model, integrating the TROPOspheric Monitoring Instrument (TROPOMI) observations offering enhanced spatiotemporal tropospheric NO_2 quantification with surface measurements across China's national monitoring network. This data fusion yielded improved daily predictions of near-surface NO_2 concentrations from 2018–2021 (R^2 = 0.73). Wei et al. [5] integrated spatiotemporally weighted information into the missing extra-trees and deep forest models to derive daily 1 km surface NO_2 concentrations over mainland China for the period 2019–2020.

Numerous studies demonstrate meaningful progress in advancing retrieved ground-level estimations of near-surface NO_2 concentrations. Nevertheless, these studies generally focus on large-scale remote sensing estimations of near-surface NO_2 concentration, either globally or nationally, providing insufficient intra-urban resolutions to direct regional mitigation strategies. Furthermore, satellite retrieval effectiveness is hampered by China's monsoonal domains with frequent meteorological obscurations such as clouds and rain, resulting in significant spatial gaps [18,23,24], further limiting the process of local governments in refining air quality management.

Against this backdrop, this study takes Chengdu, a city in the subtropical region of China, as a case study. While previous studies have focused on large-scale regional estimation of near-surface NO_2 [2,3,8], our work distinguishes itself by demonstrating an adaptive and flexible integrated approach to derive high-resolution intra-urban NO_2

maps that are crucial for informing locally tailored air quality strategies. While spatial–temporal patterns from Empirical Orthogonal Function (EOF) estimate the missing data, the dense ground observations help to provide a reliability constraint on the interpolated values. Due to its unbiased nature and robustness, combining an EOF field with ground observations allows for an estimate of the concentration fields to be made even though there are limitations on remotely sensed data due to extreme meteorological conditions including, clouds and rain [25]. As this investigation incorporates EOF interpolation with dense urban atmospheric observation station data to reconstruct continuous daily tropospheric NO_2 column concentration data, it leverages the strengths of both techniques synergistically. By comparison, traditional and widely used approaches, including LUR and GTWR models, heavily rely upon ground observation and surrounding nearby regions, which are well represented by their observations [14,18]. Hence, they do not allow direct or unbiased ways to interpolate the data into regions that are different and obscured from remotely sensed observations. Additionally, in areas with sparse or unevenly distributed monitoring sites, these traditional modeling approaches tend to suffer from spatial sampling biases.

The integrated approach used in this work bypasses these issues by using satellite data as a starting point, applying pattern recognition both spatially and temporally, and constraining these patterns using available dense networks of ground-based data when and where it is available. This proves advantageous when compared with modern gap-filling methods like EOF combined with ensemble empirical mode decomposition (EEMD), which, while powerful, may lead to biases since these data-driven techniques solely rely on the patterns obtained from the original satellite data, which may allow errors to propagate and are too computationally expensive to readily update. Additionally, the assimilation of ancillary geographic data such as meteorological and demographic information facilitates the comparative assessment of multiple predictive models for ground-level NO_2 concentrations. The high-resolution mapping methodology demonstrated in this representative subtropical case study has broader applicability as a template for urban air quality characterization in other meteorologically obscured regions globally, facilitating more precise environmental regulations and public health exposure assessments. Thus, the result aims to support the formulation of targeted air quality management policies and environmental regulations for local governments.

2. Data and Methods

2.1. Data Source

To estimate near-surface NO_2 concentrations, TROPOMI Tropospheric NO_2 data, ground monitoring station NO_2 data, and a wide range of auxiliary data known to have a physical or geographical connection with surface NO_2 concentrations were selected in this study. The auxiliary data include meteorological conditions, land cover type, topography, population distribution density, and vegetation productivity. All of the different types of data were resampled to a common spatial grid with a spatial resolution of $0.05^\circ \times 0.05^\circ$ using a bilinear interpolation algorithm. Subsequently, all datasets were projected in an Albers equal-area projection (Table 1).

Table 1. Summary of the datasets used in this study.

Name	Elements/Abbreviation	Spatial Resolutions	Temporal Resolutions	Source
Ground Monitoring Station data	NO_2	--	Hourly	CNEMC
TROPOMI	NO_2	3.5×5.5 km	Daily	Sentinel-5p
Meteorological data	U_{10}, V_{10}, T_{2m} SP, BLH, TP	$0.25^\circ \times 0.25^\circ$	Hourly	ERA5
Population data	POP	$0.01^\circ \times 0.01^\circ$	Yearly	WorldPop
Digital elevation data	DEM	30×30 m	--	GSCLOUD
Land use data	LCT	$0.05^\circ \times 0.05^\circ$	Yearly	Globeland
NDVI data	NDVI	$0.01^\circ \times 0.01^\circ$	Yearly	RESDC

2.1.1. TROPOMI Tropospheric NO_2 Data

NO_2 measurements from TROPOMI are widely utilized for estimating near-surface concentrations [26,27]. The TROPOMI is an advanced spectrometer aboard the Sentinel-5 Precursor satellite, launched in October 2017 by the European Space Agency (ESA) [28]. It measures spectra across the ultraviolet–visible (270–500 nm), near-infrared (710–770 nm), and short-wave infrared (2314–2382 nm) ranges [29]. TROPOMI facilitates global monitoring of gaseous pollutants like NO_2, which is particularly beneficial in regions with limited ground-level air quality monitoring stations. TROPOMI offers data on the tropospheric vertical column density with a spatial resolution of 3.5 km × 7 km, enhanced to 5.5 km × 3.5 km post-August 2019. Such high spatial resolution enables detailed analysis of local near-surface NO_2 distributions. Users can access imaging data for a specific region within four hours of satellite scanning, enabling timely analysis of pollutant concentration distribution characteristics. This study utilizes the TROPOMI Level-2 offline products (S5P_OFFL_L2_NO_2) for the year 2021 across China. These products are filtered to ensure data quality (qa_value > 0.75).

2.1.2. Ground Monitoring Station NO_2 Data

Figure 1 shows the spatial distribution of Chengdu's air quality monitoring stations, comprising 14 national air quality monitoring stations, 22 provincial air quality monitoring stations, and 19 municipal air quality monitoring stations. National air quality monitoring stations are unevenly distributed, primarily clustered in densely populated urban areas like the Jinjiang District. In contrast, provincial and municipal air quality monitoring stations are distributed across every county or district. The provincial and municipal air quality monitoring stations in Chengdu report daily 24-h averages for eight air pollutants, including SO_2, NO_2, and PM_{10}, along with the air quality index (AQI) and air quality levels, in accordance with the "Environmental Air Quality Standards" (GB3095-2012) [17]. The daily 24-h average concentrations of NO_2 for the year 2021 were used as the ground-level daily values for modeling.

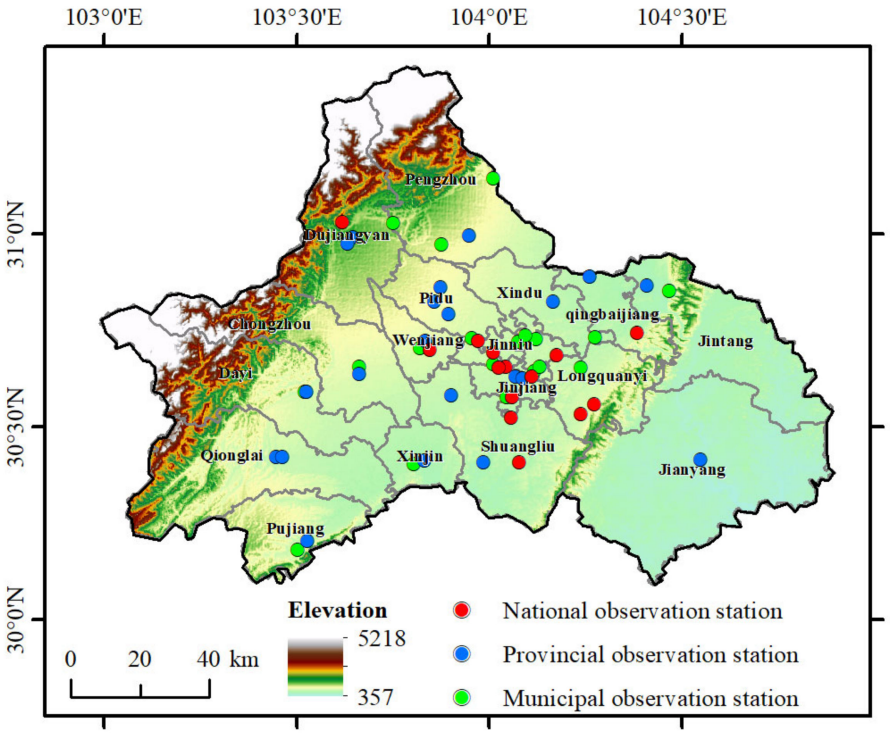

Figure 1. Spatial distribution of ground-based air quality monitoring stations in Chengdu City.

2.1.3. Auxiliary Data

Incorporating meteorological data, land cover type, topography, population distribution, and other geographic covariates into modeling can enhance the accuracy of NO_2 concentration estimates [3,5]. Meteorological variables have been proven to have significant and diverse impacts on air pollutants such as NO_2 and $PM_{2.5}$ [30]. The hourly meteorological dataset from ECMWF's ERA5, encompassing data from 1979 to the present, provides high-quality, high-resolution, and global coverage of various elements, including temperature, wind speed and direction, precipitation, and humidity [31]. According to findings of previous studies [3,5], this study selected six meteorological variables, including temperature at 10 m above the sea surface (T_{2m}), u-component and v-component of winds at 10 m above the sea surface (V_{10} and U_{10}), boundary layer height (BLH), total precipitation (TP), and surface pressure (SP) for modeling.

Additionally, population distribution density (POP) and land cover type (LCP) are closely related to the amounts of NO_2 emissions. Furthermore, topography, such as the digital elevation model (DEM), and vegetation factors, such as the normalized difference vegetation index (NDVI), indirectly influence the rates of uptake and transport of NO_2 [3,32]. These geographic factors were incorporated as auxiliary variables to model NO_2 concentrations. Population distribution density was obtained from the 2020 gridded population distribution data (WorldPop, https://wopr.worldpop.org/, accessed on 1 May 2024). DEM data was downloaded from the Geospatial Data Cloud (GSCLOUD). Land cover type data was acquired from Global Land Cover Data (Globeland, http://www.geodata.cn, accessed on 1 May 2024), covering types like cropland, water bodies, wetlands, and forests [33]. NDVI data is derived from the Annual Vegetation Index spatial grid distribution dataset by the Resource and Environmental Science and Data Center of the Chinese Academy of Sciences (RESDC, http://www.geodata.cn, accessed on 1 May 2024).

2.2. Methodology

The remote sensing estimation of near-surface NO_2 concentration using the integration of EOF interpolation, the XGBoost algorithm, and dense urban atmospheric observation station data consists of three principal technical approaches used in tandem, as illustrated in Figure 2. The first approach involves data collection, where auxiliary data such as TROPOMI satellite tropospheric NO_2 column concentration data, population grids, and ERA5 meteorological data are selected as independent variables, while ground-based monitoring of NO_2 concentration data at the national, provincial, and municipal stations is used as the dependent variables. The second approach is the derivation of the model and its validation. To accomplish this, first, the data was filtered and underwent quality control. Next, the DINEOF method was used to reconstruct gaps within the tropospheric NO_2 column concentration data from TROPOMI, with a particular interest in areas with high cloud cover. Standard matching was then performed on the temporal and spatial dimensions of all datasets. Finally, in this step, the XGBoost algorithm was employed to train and optimize the datasets, with model validation performed using metrics such as R^2 and *RMSE* derived from ten-fold cross-validation. The third component approach involves employing the just-formed model to compute estimations, thereby mapping near-surface spatial–temporal NO_2 concentration at $0.05^\circ \times 0.05^\circ$. Comparisons of predicted and mapped NO_2 are made with published datasets, including differences between normal days and holidays as well as between urban and rural areas.

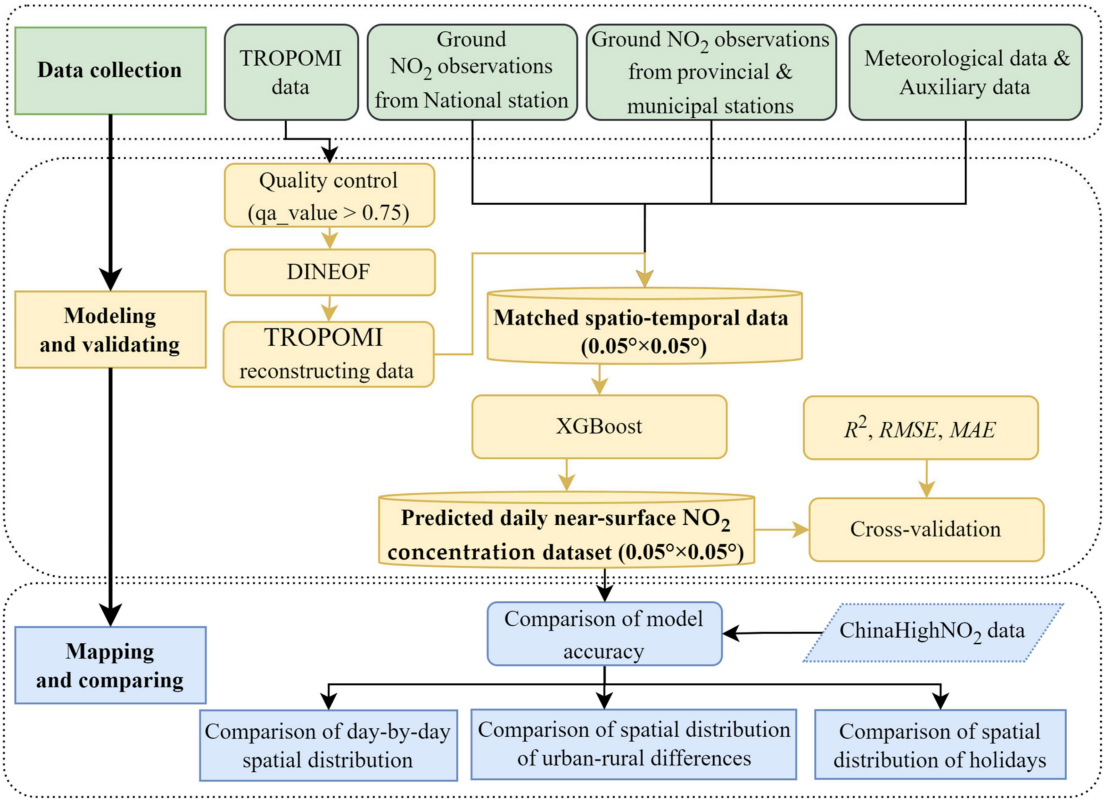

Figure 2. Flowchart of the methodology.

2.2.1. Reconstruction of Missing Remote Sensing Data

EOF analysis is widely employed to objectively quantify signals in space and time, which account for the majority of the variability of an observed dataset [25]. Within the context of this work, it is specifically used to fill gaps in long-term satellite remote sensing datasets, with an emphasis on obstructions like cloud cover [34]. This method was further developed into the Data-Interpolating Empirical Orthogonal Functions (DINEOF) by Alvera-Azcarate et al. [35]. DINEOF does not require a priori knowledge and is computationally efficient, making it well suited for reconstructing large-area remote sensing data. It is predicted that all EOF models must conduct cross-validation to ensure that the mathematical signal and the physical signal show consistency [36]. This specific approach herein utilizes iterative Singular Value Decomposition (SVD) to decompose and synthesize two data variables simultaneously, thereby obtaining an optimal set of EOF modes to reconstruct the missing NO_2 column data [37]. The new TROPOMI products (S5P_HighCoverage_NO_2) were reconstructed based on the underlying TROPOMI Level-2 offline product (S5P_OFFL_L2_NO_2).

The DINEOF used to reconstruct the TROPOMI tropospheric NO_2 data operates on a Windows platform, with Python scripts calling the dineof executable. The specific parameter settings are as follows: Firstly, the maximum number of EOF modes calculated is set to 20, while the minimum is set to 1. The size of the Krylov subspace is then set, which must be at least greater than the maximum number of EOF modes plus five, and less than or equal to the size of the data in time, with Krylov = 25 set accordingly. Additionally, the maximum number of iterations for each EOF calculation is set to 300. The iteration stops when the ratio of continuous reconstruction (RMS) to existing data (STD) drops below a threshold of 1×10^{-3}. The convergence threshold for the Lanczos method is set to 1×10^{-8} the core of the Lanczos algorithm involves using a tridiagonal matrix to find all eigenvalues within the Krylov subspace. To prevent excessive iterations, this Lanczos convergence threshold is crucial.

2.2.2. Remote Sensing Estimation of Near-Surface NO_2 Concentration

After correcting for vertical sensitivity, remote sensing measurements provide an integrated tropospheric NO_2 column amount [38], which has become one of the most effective methods for estimating NO_2 concentrations. Increasingly, studies have observed a certain correlation between tropospheric NO_2 column concentrations and ground-monitored NO_2 concentrations. However, various factors, such as traffic and industrial emissions sinks like chemical reactions, atmospheric mixing processes, and co-emitted heat, can cause significant changes in NO_2 concentrations with altitude [27]. Additionally, atmospheric conditions significantly affect the diffusion, transport, and chemical transformation of NO_2 [39]. Therefore, the relationship between tropospheric NO_2 column concentration data and near-surface NO_2 concentration is characterized by a complex nonlinearity that is hard to delineate through simplistic mathematical models [40].

Machine learning shows superior efficacy in handling these multi-factor nonlinear relationships [41]. After comparing performances between Random Forest (RF) and Extreme Gradient Boosting (XGBoost) model, XGBoost is selected for modeling and estimating near-surface NO_2 concentrations. XGBoost is an advanced ensemble machine learning algorithm that refines traditional gradient boosting by introducing robust regularization [42]. XGBoost builds upon the foundation of Gradient Boosted Regression Trees (GBRT) by incorporating advanced regularizations. Unlike GBRT, XGBoost includes both L1 and L2 regularization terms, which help reduce model overfitting and improve model generalizability. XGBoost differentiates itself by utilizing the second-order gradient, leveraging the Taylor expansion of the loss function's second derivative for more accurate approximation and effective weight updates. This method handles missing data intrinsically through a sparsity-aware split finding algorithm, facilitating optimal imputation during training. Compatible with single-machine and distributed computing frameworks, XGBoost is both scalable and efficient, suited for varied computational environments.

Recently, XGBoost has been successfully applied in remote sensing to inverse model atmospheric pollutants, demonstrating significant capability in analyzing complex environmental data sets [40–42]. Its robust predictive performance makes it ideal for such applications, combining numerous weak learners into a powerful aggregated model that provides superior predictive accuracy across various domains. The XGBoost algorithm process primarily consists of two major parts: the first part involves constructing weak learners, and the second part entails aggregating multiple weak learners to form a strong learner.

2.2.3. Experimental Grouping

To evaluate the potential of near-surface NO_2 concentration estimations by integrating reconstructed TROPOMI tropospheric NO_2 column data and denser municipal monitoring station data, this study has constructed three sets of estimation training datasets, as described in Table 2. To evaluate the above effectiveness, it is important to note that the sources and processing procedures for auxiliary covariates within these three training datasets are the same, including meteorological condition, land cover type, topography, population distribution density, and vegetation productivity.

Table 2. Experimental grouping in this study.

Experimental Grouping	Experimental Areas	TROPOMI Data	Ground NO_2 Monitoring Station Data
Group A	China	S5P_OFFL_L2_NO_2 (No reconstruction)	National stations
Group B	Chengdu	S5P_OFFL_L2_NO_2 (No reconstruction)	National, provincial, and municipal stations
Group C	Chengdu	S5P_HighCoverage_NO_2 (Reconstructing data)	National, provincial, and municipal stations

This study also conducted a comparative analysis to evaluate the enhanced capability of estimating NO_2 concentrations using EOF interpolation combined with data from densely located urban atmospheric observation stations. It compared these results with the high-resolution and high-quality ground-level NO_2 dataset for China (ChinaHighNO_2) [3,5], which covers daily, seamless, and nationwide ground-level NO_2 data with a spatial resolution of 1 × 1 km from 2008 to 2022. This dataset was modeled by multiple sources of datasets, including ground-based NO_2 observations, satellite remote sensing products, atmospheric reanalysis, and model simulations. The products from the year 2021 were selected and then resampled to a spatial resolution of 0.05° × 0.05° using the bilinear algorithm.

3. Results and Discussion

3.1. Results of Reconstructing Remote Sensing Products

To understand the differences in the percentage of annual coverage of valid data before and after the reconstruction of missing data of TROPOMI NO_2 column concentration in Chengdu City, this study analyzes and evaluates the coverage by comparing the percentage of valid data days in each grid cell to the total number of days in 2021 (365 days). Figure 3a shows that the original TROPOMI NO_2 column concentration data coverage is relatively low, with an average value of 29%. Following the DINEOF reconstruction, as shown in Figure 3b, the average coverage rate of the TROPOMI tropospheric NO_2 column concentration data is elevated to 99.2%. Figure 3c is obtained by computing the difference between the coverage in each effective grid cell after reconstruction and before, which spatially reveals that the increase in coverage in the northwestern regions of Chengdu City is higher than in other regions.

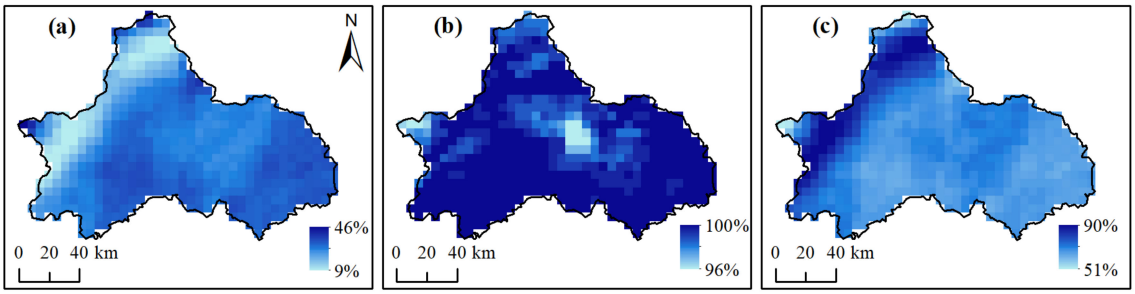

Figure 3. Comparison of the spatial pattern of annual coverage of available data before (**a**) and after (**b**) the reconstruction of tropospheric NO_2 column concentration in TROPOMI in 2021. (**c**) denotes the difference in the coverage of available data between before and after the reconstruction.

3.2. Evaluation of Model Performance

We employed the XGBoost model to estimate the near-surface NO_2 concentration, evaluated the performance of the model using high-density ground-based NO_2 measurements, and reconstructed the TROPOMI dataset through mathematical indexes, as shown in Table 3 and Figure 4. The model performance of Group B (R^2 = 0.87, *RMSE* = 5.80 μg/m^3) surpassed that of Group A (R^2 = 0.812, *RMSE* = 7.14 μg/m^3), signifying that incorporating denser monitoring station data can enhance the precision of the XGBoost model in estimating near-surface NO_2 concentration. Moreover, the model performance of Group C (R^2 = 0.87, *RMSE* = 5.36 μg/m^3) uses reconstructed satellite data. The performance of Group C not only exceeds that of Groups A and B, but also shows better performance than that of Zhan and Luo et al. [32] (R^2 = 0.62, *RMSE* = 13.3 μg/m^3) using the (RF-STK) to estimate regional near-surface daily NO_2 concentration in China, and that of Chi and Fan et al. [22] (R^2 = 0.73, *RMSE* = 5.63 μg/m^3) in estimating near-surface NO_2 concentration in China.

Table 3. The performance of the XGBoost model in each group of experiments.

Experimental Grouping		R^2	RMSE	MAE
Group A	Test Set	0.81	7.14	5.25
	Training set	0.84	6.60	4.87
Group B	Test Set	0.87	5.80	4.33
	Training set	0.98	2.12	1.58
Group C	Test Set	0.87	5.36	3.96
	Training set	0.94	3.57	2.68

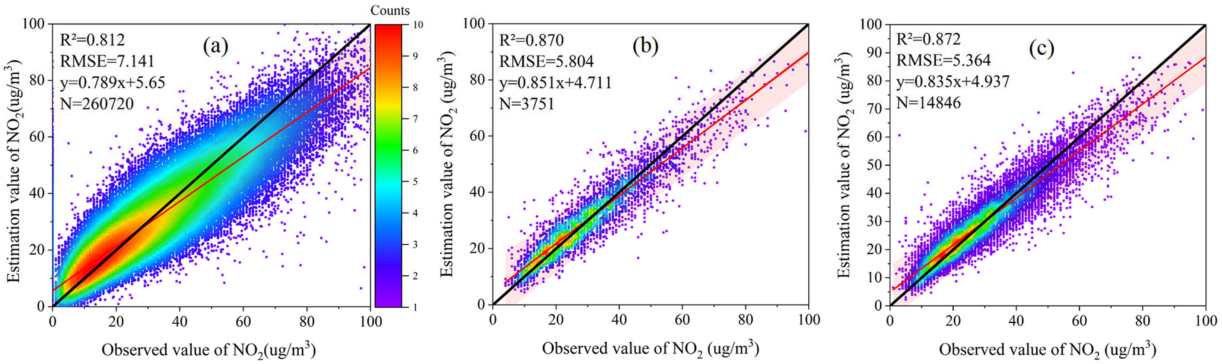

Figure 4. The performance of the XGBoost model for each group of experiments. (**a**–**c**) denote the results of Group A, Group B, and Group C.

3.3. Analysis of Mapping Results

3.3.1. Comparison of Daily Spatial Distribution of Near-Surface NO_2 Concentration

The near-surface NO_2 concentration estimated with dense air quality observed station data (including province and municipal monitoring stations) is closer to the measured NO_2 concentration of the ground-based stations when there are no missing values from the satellites, as shown in Figure 5a–f. Substantial data gaps in Chengdu were observed due to satellite omissions caused by cloud and fog cover during 21–23 March. Spatially, valid data was available in the middle part of Chengdu City on 21 March, and in small areas of the northern and western regions on 23 March. Consequently, significant missing values were evident in the spatial distribution of near-surface NO_2 estimates for Groups A and B. Compared with Groups A and B, Group C has successfully filled the areas with missing measurements (Figure 5g–i). Group C achieved a high level of fit, in some cases consistent with ChinaHighNO_2 produced by Wei et al. [3], and in some cases even closer to the observed surface monitoring stations. Group C results in particular show a better match under medium and medium–high pollution conditions (Figure 5g–o).

3.3.2. Comparison of Urban–Rural Spatial Distribution of Near-Surface NO_2 Concentration

In urban centers of Chengdu City, such as Jinniu District, Wenjiang District, and Wuhou District, the near-surface NO_2 concentrations are maintained at relatively high levels. Figure 5 indicates that the utilization of reconstructed satellite data in these regions has exacerbated the underestimation of NO_2 concentration. This may be attributed to the city's consistent presence in an area of high NO_2 values and the potential underestimation by TROPOMI tropospheric data in these highly polluted regions of Chengdu City. The disparity in the underestimation phenomena for estimated near-surface NO_2 concentration is relatively minor in suburban counties like Pujiang County and Dayi County.

While both our estimations and ChinaHighNO_2 effectively mirror the spatial distribution of near-surface NO_2 concentration across Chengdu City, our version does better at not overestimating the highest pixels while capturing a more realistic distribution of medium pollution level pixels. The ChinaHighNO_2 presents a nuanced portrayal of cartographic distinctions due to its superior spatial resolution of 1 × 1 km. However, the estimations from this study show superior alignment with actual monitoring station data in certain

areas (Figure 5g–o), likely due to the fact that there is less overfitting when transforming the data from the observed grid to the 0.05° × 0.05° grid, as compared to a 1 × 1 km grid, which is further from the originally observed data.

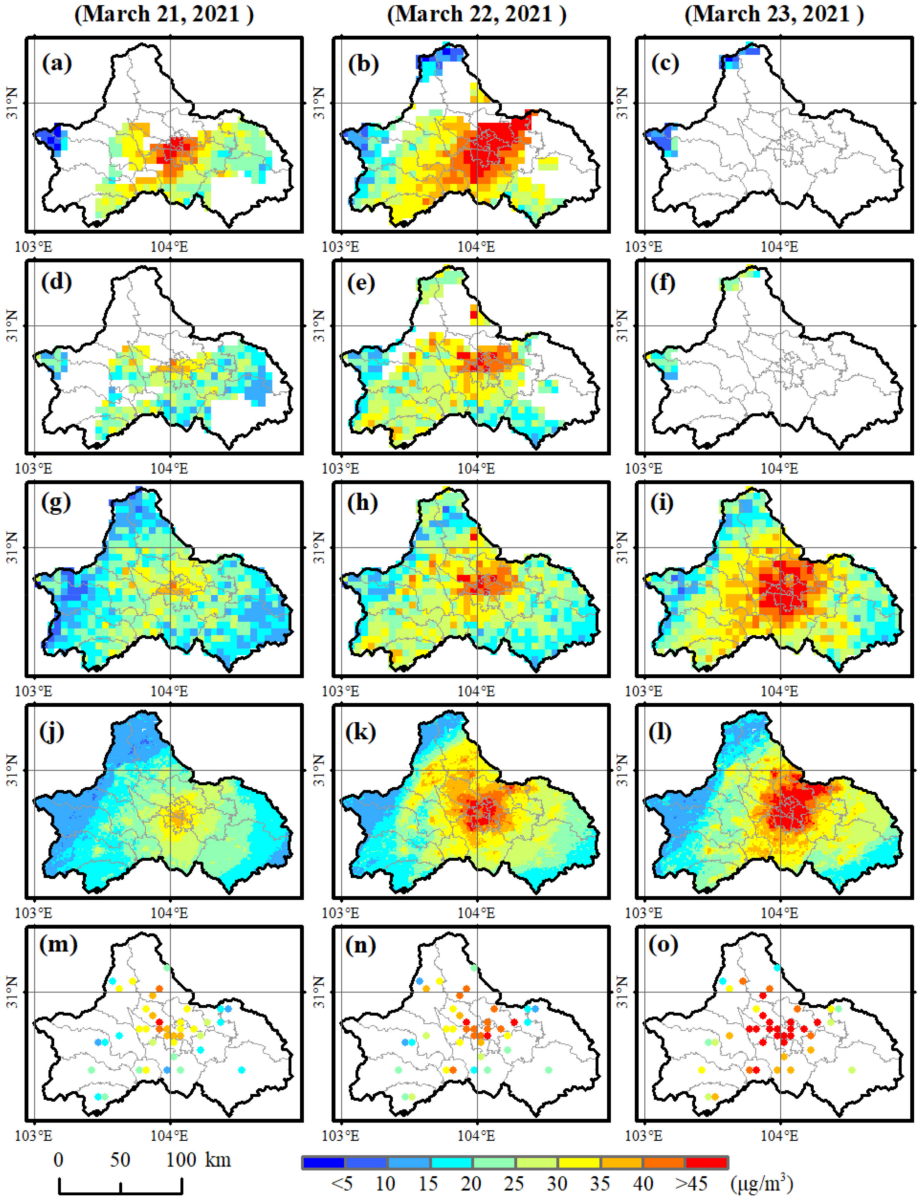

Figure 5. Comparison of near-ground NO_2 concentration estimation on 21–23 March for each group as well as the ChinaHighNO_2 produced by Wei et al. [3]. (**a**–**c**) belong to Group A; (**d**–**f**) belong to Group B; (**g**–**i**) belong to Group C; (**j**–**l**) belong to the ChinaHighNO_2 data; and (**m**–**o**) belong to the observed value of ground station. (**a**,**d**,**g**,**j**,**m**), (**b**,**e**,**h**,**k**,**n**), and (**c**,**f**,**i**,**l**,**o**) denote NO_2 concentration estimation on 21–23 March, respectively.

The performance is particularly improved in areas with significant urban–rural transition zones, such as Wenjiang District and Chongzhou City, where there is a large disparity between high and low values. This improvement is largely attributable to the inclusion of a denser network of monitoring stations in the estimation process. As shown in Figure 1, national air quality monitoring stations are primarily clustered in densely populated urban areas, such as Jinjiang District, while provincial and municipal air quality monitoring stations are distributed across every county or district.

3.3.3. Comparison of Holiday Spatial Distribution of Near-Surface NO_2 Concentration

To demonstrate the importance of using reconstructed satellite data and integrating denser ground monitoring station data in estimating near-surface NO_2 concentration, this study also conducted a comparative analysis of the spatial distribution data of near-surface NO_2 concentration and the measured ground station data during the Spring Festival holiday period of 2021 (Figure 6). As shown in Figure 6, Groups A and B, which did not incorporate reconstructed data, display evident gaps and relatively poor spatial continuity. At the relatively high NO_2 concentration on 10 February and 11 February 2021, the overestimation is more pronounced in the high-value areas in the city center. In contrast, the estimation results of Group C align closely with the actual measured NO_2 concentration distribution, showcasing enhanced image continuity and smoothness. Additionally, there was a notable decrease in NO_2 concentration in Chengdu around Chinese New Year's Eve (10 February 2021) and Spring Festival (11 February 2021), possibly attributed to the heightened emissions just before and rapid decrease in industrial activities and transportation during the festive period, consistent with Li et al. [43]. During this special time, the NO_2 concentration estimated in this study is close to that from monitoring stations in the central and southern regions of Chengdu on 11 February, with values of around 35 μg/m^3, including the Jinniu District and Shuangliu District, while the ChinaHighNO_2 overestimated NO_2 concentration in these areas, with a value of around 45 μg/m^3 (Figures 6j,n,r and S1).

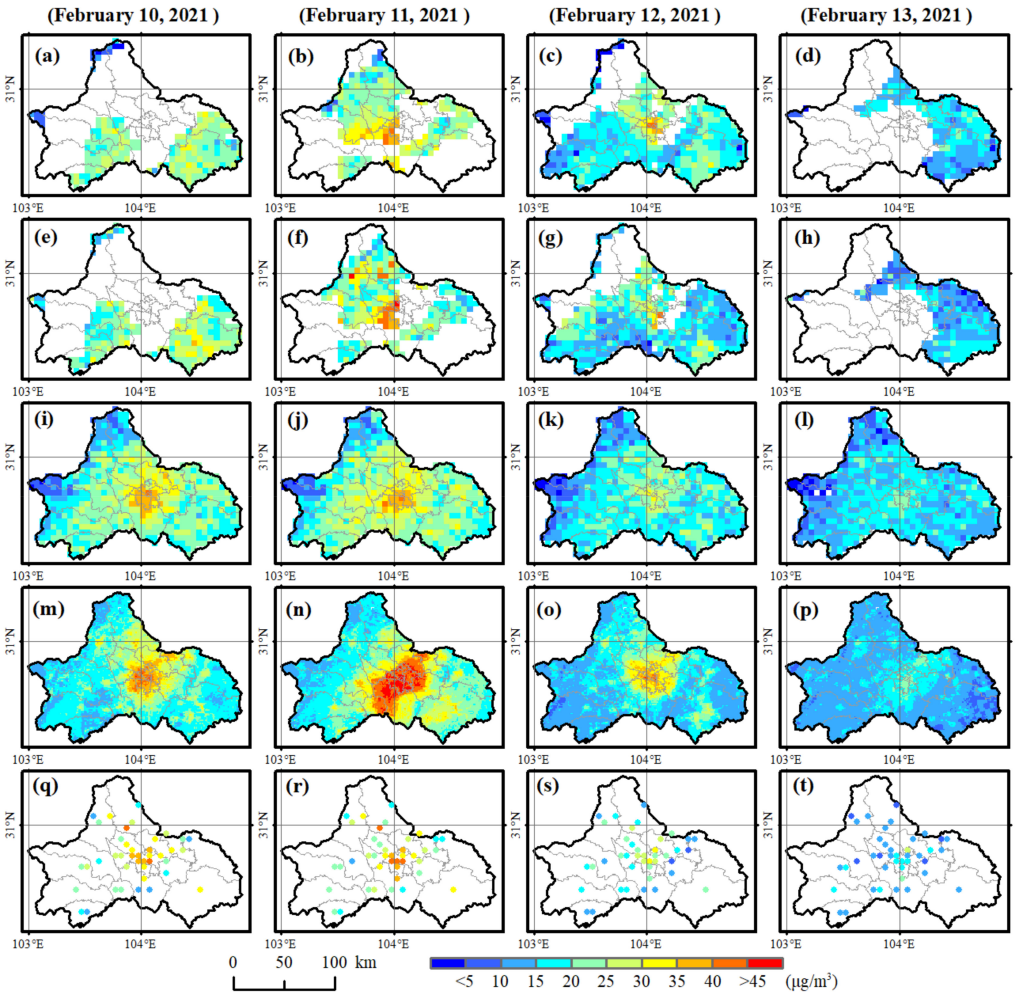

Figure 6. Comparison of near-ground NO_2 concentration estimation during the Spring Festival holiday period in 2021 (10–13 February) for each group as well as the ChinaHighNO_2 produced by Wei et al. [3]. (**a**–**d**) belong to Group A; (**e**–**h**) belong to Group B; (**i**–**l**) belong to Group C;

(**m**–**p**) belong to the ChinaHighNO_2 data; and (**q**–**t**) belong to the observed value of ground station. (**a**,**e**,**i**,**m**,**q**), (**b**,**f**,**j**,**n**,**r**), (**c**,**g**,**k**,**o**,**s**), and (**d**,**h**,**l**,**p**,**t**) denote near-ground NO_2 concentration estimation on 10–13 February, respectively.

4. Conclusions

This study presents a novel approach that integrates EOF interpolation, the XGBoost algorithm, and dense urban atmospheric observation station data to estimate near-surface NO_2 concentrations in regions plagued by frequent cloud cover and precipitation, using Chengdu as a case study area. The key conclusions derived from the research are as follows:

(1) The DINEOF method was successfully employed to reconstruct the tropospheric NO_2 column concentration data from TROPOMI for Chengdu. The reconstruction process increased the original data coverage from a mere 29% to an impressive 99.2% by optimally retaining EOF modes and iterating the process. Spatially, the reconstructed data closely matched the distribution of observed data, effectively filling in gaps caused by clouds and fog, thereby providing a comprehensive foundation for subsequent refined NO_2 estimation.

(2) Based on the findings in (1), comparative experiments were conducted using the XGBoost machine learning model with various training and estimation datasets. The results demonstrated that the model accuracy achieved by incorporating reconstructed satellite data and dense province and municipal ground station measurements (R^2 = 0.87, RMSE = 5.364 $\mu g/m^3$) was superior to models relying solely on national control site data and original satellite data (R^2 = 0.81, $RMSE$ = 7.14 $\mu g/m^3$).

(3) In analyzing the daily, holiday-based, and urban–rural spatial distribution variability of near-surface NO_2 concentration, the integration of dense ground monitoring station data and reconstructed satellite data proved instrumental. This approach effectively reduced overestimation in low-value locations, enhancing the continuity and smoothness of the spatial distribution. Consequently, the resulting estimates provided a more realistic representation of the fluctuations in near-surface NO_2 concentrations.

In conclusion, this study presents a robust methodology for estimating near-surface NO_2 concentrations in areas prone to frequent cloud cover and precipitation by synergistically combining reconstructed satellite data, ground-based measurements, and advanced machine learning techniques. The findings underscore the importance of data reconstruction and multi-source data integration in overcoming the limitations of traditional remote sensing approaches, paving the way for improved air quality monitoring and management strategies.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/rs16101785/s1. Figure S1: Comparison of newly near-ground NO_2 concentration estimation on 11 February with the ChinaHighNO_2 produced by Wei et al. [3]. (a,b) denote our new results and ChinaHighNO_2, respectively.

Author Contributions: Conceptualization and methodology, F.D., Y.C. and K.Q.; investigation and data curation, F.D., Y.C., L.L., X.C. and W.L.; writing—original draft preparation, F.D. and Y.C.; writing—review and editing, F.D., Y.C., P.T. and K.Q.; supervision, K.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This project was financially supported by the National Natural Science Foundation of China (Grant Number 42375125).

Data Availability Statement: Data code will be made available on reasonable request.

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

References

1. Solomon, S.D.; Qin, D.; Manning, M.; Chen, Z.; Miller, H.L. Climate Change 2007: The physical science basis. Working Group I contribution to the fourth assessment report of the IPCC. In *Intergovernmental Panel on Climate Change Climate Change*; Cambridge University Press: Cambridge, UK, 2007.
2. De Craemer, S.; Vercauteren, J.; Fierens, F.; Lefebvre, W.; Meysman, F.J.R. Using Large-Scale NO_2 Data from Citizen Science for Air-Quality Compliance and Policy Support. *Environ. Sci. Technol.* **2020**, *54*, 11070–11078. [CrossRef] [PubMed]
3. Wei, J.; Li, Z.; Wang, J.; Li, C.; Gupta, P.; Cribb, M. Ground-level gaseous pollutants (NO_2, SO_2, and CO) in China: Daily seamless mapping and spatiotemporal variations. *Atmos. Chem. Phys.* **2023**, *23*, 1511–1532. [CrossRef]
4. He, T.; Tang, Y.; Cao, R.; Xia, N.; Li, B.; Du, E. Distinct urban-rural gradients of air NO_2 and SO_2 concentrations in response to emission reductions during 2015–2022 in Beijing, China. *Environ. Pollut.* **2023**, *333*, 122021. [CrossRef] [PubMed]
5. Wei, J.; Liu, S.; Li, Z.; Liu, C.; Qin, K.; Liu, X.; Pinker, R.T.; Dickerson, R.R.; Lin, J.; Boersma, K.F.; et al. Ground-Level NO_2 Surveillance from Space Across China for High Resolution Using Interpretable Spatiotemporally Weighted Artificial Intelligence. *Environ. Sci. Technol.* **2022**, *56*, 9988–9998. [CrossRef] [PubMed]
6. Huang, S.; Li, H.; Wang, M.; Qian, Y.; Steenland, K.; Caudle, W.M.; Liu, Y.; Sarnat, J.; Papatheodorou, S.; Shi, L. Long-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. *Sci. Total Environ.* **2021**, *776*, 145968. [CrossRef] [PubMed]
7. Wang, M.; Li, H.; Huang, S.; Qian, Y.; Steenland, K.; Xie, Y.; Papatheodorou, S.; Shi, L. Short-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. *Environ. Res.* **2021**, *202*, 111766. [CrossRef] [PubMed]
8. Wu, S.; Huang, B.; Wang, J.; He, L.; Wang, Z.; Yan, Z.; Lao, X.; Zhang, F.; Liu, R.; Du, Z. Spatiotemporal mapping and assessment of daily ground NO_2 concentrations in China using high-resolution TROPOMI retrievals. *Environ. Pollut.* **2021**, *273*, 116456. [CrossRef] [PubMed]
9. Kong, H.; Lin, J.; Chen, L.; Zhang, Y.; Yan, Y.; Liu, M.; Ni, R.; Liu, Z.; Weng, H. Considerable Unaccounted Local Sources of NOx Emissions in China Revealed from Satellite. *Environ. Sci. Technol.* **2022**, *56*, 7131–7142. [CrossRef] [PubMed]
10. Tang, D.; Zhan, Y.; Yang, F. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmos. Res.* **2024**, *300*, 107261. [CrossRef]
11. Martin, R.V.; Chance, K.; Jacob, D.J.; Kurosu, T.P.; Spurr, R.J.D.; Bucsela, E.; Gleason, J.F.; Palmer, P.I.; Bey, I.; Fiore, A.M.; et al. An improved retrieval of tropospheric nitrogen dioxide from GOME. *J. Geophys. Res. Atmos.* **2002**, *107*, ACH 9-1–ACH 9-21. [CrossRef]
12. Lamsal, L.N.; Martin, R.V.; van Donkelaar, A.; 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.* **2008**, *113*, D16308. [CrossRef]
13. Hystad, P.; Setton, E.; Cervantes, A.; Poplawski, K.; Deschenes, S.; Brauer, M.; van Donkelaar, A.; Lamsal, L.; Martin, R.; Jerrett, M.; et al. Creating national air pollution models for population exposure assessment in Canada. *Environ. Health Perspect.* **2011**, *119*, 1123–1129. [CrossRef] [PubMed]
14. Knibbs, L.D.; Hewson, M.G.; Bechle, M.J.; Marshall, J.D.; Barnett, A.G. A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environ. Res.* **2014**, *135*, 204–211. [CrossRef] [PubMed]
15. Gulliver, J.; de Hoogh, K.; Hansell, A.; Vienneau, D. Development and Back-Extrapolation of NO_2 Land Use Regression Models for Historic Exposure Assessment in Great Britain. *Environ. Sci. Technol.* **2013**, *47*, 7804–7811. [CrossRef] [PubMed]
16. Larkin, A.; Geddes, J.A.; Martin, R.V.; Xiao, Q.; Liu, Y.; Marshall, J.D.; Brauer, M.; Hystad, P. Global Land Use Regression Model for Nitrogen Dioxide Air Pollution. *Environ. Sci. Technol.* **2017**, *51*, 6957–6964. [CrossRef] [PubMed]
17. *GB 3095-2012*; Environmental Air Quality Standards. Ministry of Environmental Protection of PRC; General Administration of Quality Supervision, Inspection and Quarantine of PRC: Beijing, China, 2012.
18. 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]
19. Wang, Y.; Yuan, Q.; Li, T.; Zhu, L.; Zhang, L. Estimating daily full-coverage near surface O_3, CO, and NO_2 concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 311–325. [CrossRef]
20. Araki, S.; Shima, M.; Yamamoto, K. Spatiotemporal land use random forest model for estimating metropolitan NO_2 exposure in Japan. *Sci. Total Environ.* **2018**, *634*, 1269–1277. [CrossRef]
21. You, J.; Zou, B.; Zhao, X.; Xu, S.; He, R. Estimating ground-level NO_2 concentrations across mainland China using random forests regression modeling. *Chin. J. Environ. Sci.* **2019**, *39*, 969–979.
22. Chi, Y.; Fan, M.; Zhao, C.; Yang, Y.; Fan, H.; Yang, X.; Yang, J.; Tao, J. Machine learning-based estimation of ground-level NO_2 concentrations over China. *Sci. Total Environ.* **2022**, *807*, 150721. [CrossRef]
23. Kim, M.; Brunner, D.; Kuhlmann, G. Importance of satellite observations for high-resolution mapping of near-surface NO_2 by machine learning. *Remote Sens. Environ.* **2021**, *264*, 112573. [CrossRef]
24. Li, L.; Wu, J. Spatiotemporal estimation of satellite-borne and ground-level NO_2 using full residual deep networks. *Remote Sens. Environ.* **2021**, *254*, 112257. [CrossRef]
25. Cohen, J.B. Quantifying the occurrence and magnitude of the Southeast Asian fire climatology. *Environ. Res. Lett.* **2014**, *9*, 114018. [CrossRef]

26. 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. [CrossRef]
27. Fan, C.; Li, Z.; Li, Y.; Dong, J.; van der A, R.; de Leeuw, G. Variability of NO_2 concentrations over China and effect on air quality derived from satellite and ground-based observations. *Atmos. Chem. Phys.* **2021**, *21*, 7723–7748. [CrossRef]
28. Veefkind, J.P.; Aben, I.; McMullan, K.; Förster, H.; de Vries, J.; Otter, G.; Claas, J.; Eskes, H.J.; de Haan, J.F.; Kleipool, Q.; et al. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sens. Environ.* **2012**, *120*, 70–83. [CrossRef]
29. Geffen, J.V.; Boersma, K.F.; Eskes, H.; Sneep, M.; Veefkind, J.P. S5P TROPOMI NO_2 slant column retrieval: Method, stability, uncertainties and comparisons with OMI. *Atmos. Meas. Tech.* **2020**, *13*, 1315–1335. [CrossRef]
30. Li, R.; Wang, Z.; Cui, L.; Fu, H.; Zhang, L.; Kong, L.; Chen, W.; Chen, J. Air pollution characteristics in China during 2015–2016: Spatiotemporal variations and key meteorological factors. *Sci. Total Environ.* **2019**, *648*, 902–915. [CrossRef] [PubMed]
31. Uppala, S.M.; Dee, D.; Kobayashi, S.; Berrisford, P.; Simmons, A. Towards a climate data assimilation system: Status update of ERA-Interim. *ECMWF Newsl.* **2008**, *115*, 12.
32. 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. [CrossRef] [PubMed]
33. Chen, J.; Ban, Y.; Li, S. Open access to Earth land-cover map. *Nature.* **2014**, *7523*, 434.
34. Beckers, J.M.; Rixen, M. EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *J. Atmos. Ocean. Technol.* **2003**, *12*, 1839–1856. [CrossRef]
35. Alvera-Azcárate, A.; Barth, A.; Sirjacobs, D.; Lenartz, F.; Beckers, J.M. Data Interpolating Empirical Orthogonal Functions (DINEOF): A tool for geophysical data analyses. *Mediterr. Mar. Sci.* **2011**, *12*, 5–11. [CrossRef]
36. Lin, C.; Cohen, J.B.; Wang, S.; Lan, R. Application of a combined standard deviation and mean based approach to MOPITT CO column data, and resulting improved representation of biomass burning and urban air pollution sources. *Remote Sens. Environ.* **2020**, *241*, 111720. [CrossRef]
37. Hilborn, A.; Costa, M. Applications of DINEOF to Satellite-Derived Chlorophyll-a from a Productive Coastal Region. *Remote Sens.* **2018**, *10*, 1449. [CrossRef]
38. Richter, A.; Burrows, J.P.; Nüß, H.; Granier, C.; Niemeier, U. Increase in tropospheric nitrogen dioxide over China observed from space. *Nature* **2005**, *437*, 129–132. [CrossRef] [PubMed]
39. Liu, M.; Lin, J.; Kong, H.; Boersma, K.F.; Eskes, H.; Kanaya, Y.; He, Q.; Tian, X.; Qin, K.; Xie, P.; et al. A new TROPOMI product for tropospheric NO_2 columns over East Asia with explicit aerosol corrections. *Atmos. Meas. Tech.* **2020**, *13*, 4247–4259. [CrossRef]
40. Cohen, J.B.; Prinn, R.G. Development of a fast, urban chemistry metamodel for inclusion in global models. *Atmos. Chem. Phys.* **2011**, *11*, 7629–7656. [CrossRef]
41. Li, Y.; Qin, K.; Li, D.; Wen-zhi, F.; Qin, H.E.; Abell, R. Estimation of ground-level ozone concentration based on GBRT. *China Environ. Sci.* **2020**, *40*, 997–1007.
42. Chen, T.; Guestrin, C. *XGBoost: A Scalable Tree Boosting System*; ACM: Ithaca, NY, USA, 2016; pp. 785–794.
43. Li, X.; Cohen, J.B.; Qin, K.; Geng, H.; Wu, X.; Wu, L.; Yang, C.; Zhang, R.; Zhang, L. Remotely sensed and surface measurement-derived mass-conserving inversion of daily NO_x emissions and inferred combustion technologies in energy-rich northern China. *Atmos. Chem. Phys.* **2023**, *23*, 8001–8019. [CrossRef]