frontiers | Frontiers in Environmental Science

TYPE Brief Research Report
PUBLISHED 19 July 2023
DOI 10.3389/fenvs.2023.1187592

OPEN ACCESS

EDITED BY
Maged Marghany,
Syiah Kuala University, Indonesia

REVIEWED BY
Jian Xu,
Chinese Academy of Sciences (CAS), China
Simone Lolli,
National Research Council (CNR), Italy

*CORRESPONDENCE
Truong Xuan Ngo,
truonggnx@vnu.edu.vn

RECEIVED 16 March 2023
ACCEPTED 04 July 2023
PUBLISHED 19 July 2023

CITATION
Ngo TX, Phan HDT and Nguyen TTN (2023), Development of ground-level NO_2 models in Vietnam using machine learning and satellite observations with ancillary data.
Front. Environ. Sci. 11:1187592.
doi: 10.3389/fenvs.2023.1187592

Development of ground-level NO_2 models in Vietnam using machine learning and satellite observations with ancillary data

Truong Xuan Ngo*, Hieu Dang Trung Phan and Thanh Thi Nhat Nguyen

Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

In this study, the aim was to create daily ground-level NO_2 maps for Vietnam spanning from 2019 to 2021. To achieve this, various machine learning models (including the Mixed Effect Model, Neural Network, and LightGBM) were utilized to process satellite NO_2 tropospheric columns from Ozone Monitoring Instrument (OMI) and TROPOMI, as well as meteorological and land use maps and ground measurement NO_2 data. The LightGBM model was found to be the most effective, producing results with a Pearson r of 0.77, RMSE of 7.93 μg/m^3, and Mean Relative Error (MRE) of 42.6% compared to ground truth measurements. The annual average NO_2 maps from 2019–2021 obtained by the LightGBM model for Vietnam were compared to a global product and ground stations, and it was found to have superior quality with Pearson r of 0.95, RMSE of 2.27 μg/m^3, MRE of 9.79%, based on 81 samples.

KEYWORDS

Sentinel 5p, OMI, ground-level NO_2 model, machine learning, Vietnam

1 Introduction

Air pollution poses a significant threat to the environment and human health in many countries. In Vietnam, Nitrogen dioxide (NO_2) is recognized as a particularly important air pollutant. To monitor and manage the levels of NO_2 and other harmful pollutants such as $PM_{2.5}$, PM_{10}, SO_2, and O_3, the Ministry of Natural Resources and Environment (MONRE) has implemented automatic and continuous monitoring systems. However, the current monitoring of NO_2 in Vietnam is limited due to the lack of representative monitoring stations across the country. In recent times, modeling techniques utilizing data from monitoring stations, satellite imagery (remote sensing), and auxiliary sources have gained widespread acceptance in generating spatial NO_2 information. This approach provides additional data to supplement the readings from monitoring stations, thus providing insights into the distribution of NO_2 concentrations on a larger scale, especially in regions without monitoring stations. The NO_2 satellites used for this purpose include the Ozone Monitoring Instrument (OMI), Global Ozone Monitoring Experiment–2 (GOME-2), SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY), and TROPOspheric Monitoring Instrument (TROPOMI).

Many studies have been conducted globally to map NO_2 using satellite imagery. For instance, Larkin et al. (2017) used a land use regression (LUR) model to estimate global NO_2

levels in 2011 with a resolution of 100 × 100 m. They incorporated model and satellite data/model data from SCIAMACHY, GOME-2, and GEOS Chem, as well as land cover features such as vegetation index, tree cover, traffic, *etc.*, and monitoring station data from 58 countries. The model's performance varied depending on the region, with the coefficient of determination (R^2) ranging from 0.42 in Africa to 0.67 in South America. In North America, Europe, and Asia, the R^2 value was approximately 0.52, which is consistent with the global average (0.54) (Larkin et al., 2017). To further enhance the accuracy of NO_2 mapping, a study conducted by Anenberg et al. (2022) estimated the global average annual NO_2 levels from 1990 to 2020 at a resolution of 1 × 1 km. This study used Land Use Regression (LUR) incorporating OMI NO_2 and MERRA2-reanalysis data. Results indicate that the new NO_2 concentration data is more precise than that of Larkin's study in rural areas, with a Pearson r of 0.58 and a Root mean square error (RMSE) of 2.26 (ppb) (Anenberg et al., 2022). The results of this study have important implications for public health, as they were able to estimate the NO_2-attributable pediatrics asthma incidence using the improved NO_2 concentration data. Paraschiv examined the relationship between OMI data and monitoring stations across Europe during the period of 2005–2014. Their findings indicate a Pearson r value ranging from 0.53 to 0.86 (Paraschiv et al., 2017). Hyung Joo Lee and colleagues (2014) developed a mixed-effect model (MEM) to estimate daily NO_2 concentrations in New England, United States from 2005–2010. Their model was based on various data sources, including station data, tropospheric column NO_2 (OMI), historical land use data such as population density, traffic, topography, as well as meteorological data such as temperature and wind speed. They evaluated the model using a 10-fold cross-validation (CV) method and found an R^2 value of 0.79, indicating good model performance (Lee and Koutrakis, 2014). In the mentioned studies, OMI NO_2 satellite data is commonly used to estimate NO_2 maps.

Recently, some studies have been conducted using TROPOMI satellite data (the most recently launched satellite with high resolution data) with Machine Learning models and auxiliary data to estimate ground-level pollutant concentrations (e.g., NO_2, O_3). A study by Kang et al. (2021) estimated ground-level NO_2 and O_3 with a resolution of 6 × 6 km at East Asia using NO_2 data from the TROPOMI satellite, other satellite products (Landcover, Aerosol Optical Depth - AOD, Digital Elevation Model - DEM), meteorological data from models, and auxiliary data (road density, population density). Several different machine learning models were experimented, including Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). XGBoost showed better results when estimating NO_2 with a 10-fold cross-validation R^2 of 0.7 and RMSE of 4.75 ppb. Long et al. (2022) map daily ground-level NO_2 concentrations in China at a resolution of 0.05° using machine learning models based on decision trees (Decision Tree, Gradient Boost Decision Tree, Random Forest, Extra-Trees). They found that the Extra-Trees model incorporating spatial and temporal information performed exceptionally well in estimating ground-level NO_2 concentrations, achieving a cross-validation R^2 of 0.81 and an RMSE of 3.45 μg/m^3 in test datasets (Long et al., 2022). Wang et al. (2022) used Random Forest to estimate the daily maximum 8-hour average ground-level ozone concentration at a 10 km spatial resolution in California. They utilized TROPOMI's total ozone column combined with ozone profile information retrieved by the Ozone Monitoring Instrument (OMI) and auxiliary data (meteorological, land use). Their model achieved an overall 10-fold CV R^2 of 0.84 and an RMSE of 0.0059 ppm. In another study, Grzybowski et al. (2023) employed various data sources, including Sentinel-5P, meteorological data, and other ancillary data, to estimate ground NO_2 levels in Poland. Among the methods used, the random forest (RF) model emerged as the most accurate, with mean absolute error (MAE) values of 3.4 μg/m^3 and 3.2 μg/m^3 for the hourly and weekly estimates, respectively. The corresponding mean absolute percentage error (MAPE) values were 37% and 31%, indicating relatively moderate deviations from the true values (Grzybowski et al., 2023). The tree-based model demonstrates strong estimation capabilities in air pollution estimation problems using remote sensing and auxiliary data.

Currently, there are no studies on nationwide NO_2 estimation in Vietnam utilizing satellite images and multi-source data. However, a study conducted in 2015 developed daily $PM_{2.5}$ maps for Vietnam from 2010–2014 using a multivariable regression model (Nguyen et al., 2015). Recently, a study provided a long-term daily $PM_{2.5}$ map for Vietnam from 2012–2020 using mixed effect models based on ground $PM_{2.5}$ measurements, integrated satellite Aerosol Optical Depth (AOD), meteorological and land use maps (Ngo et al., 2023). The daily mean $PM_{2.5}$ maps have high validation results with ground $PM_{2.5}$ measurements, achieving a Pearson r of 0.87, R^2 of 0.75, RMSE of 11.76 μg/m^3, and MRE of 36.57% on a total of 13,886 data samples.

This study aimed to develop daily ground-level NO_2 maps with a resolution of 1 × 1 km over Vietnam using satellite images and multi-source data from 2019–2021. The NO_2 tropospheric columns were derived from OMI and TROPOMI satellite devices, and different models such as Mixed Effect Model, Neural Network, and LightGBM were tested. Although the models are not new, this is the first study to experimentally construct a high-resolution NO_2 map for the entire territory of Vietnam based on satellite data. Various machine learning models were experimented to find the optimal model that fits the data in Vietnam. The NO_2 maps hold promise in providing useful information on NO_2 distribution across Vietnam, supporting decision-making and policies to reduce NO_2 pollution and improving public health.

2 Materials

2.1 Measurement data

The hourly ground measurements of NO_2 were collected from monitoring stations in Vietnam. Vietnam is situated in the East of the Indochina peninsula, at the heart of Southeast Asia, with its land area covering 331,236 km^2, stretching from (8°27′N, 102°8′E) to (23°23′N, 109°27′E). The country is divided into six distinct economic zones, namely, the Northern Midlands and Mountains,

Red River Delta (RRD), North Central Coast and South Central Coast, Central Highlands, South East, and Mekong River Delta (MRD) as illustrated in Supplementary Figure S1.

The Northern Center for Environmental Monitoring (NCEM), which operates under the Vietnam Administration of Environment (VEA) under MONRE, is responsible for air pollution monitoring in Vietnam. As of 2021, over 90 stations have been installed across the country, with most of them located in the Red River Delta (RRD) region. These stations measure various pollutants such as NO_2, PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, as well as meteorological variables like temperature, humidity, and wind speed. Hourly NO_2 concentration ($\mu g/m^3$) data from 74 stations were collected between 2019–2021 in this study, with poor quality data stations removed. The distribution of ground stations is illustrated in Supplementary Figure S1.

2.2 Satellite data

In order to monitor air pollution at stations on a national scale, satellite images are also used which has a larger coverage than the traditional monitoring method. The development of satellite technology can solve the problem of monitoring air pollution on a large scale. For this study, we utilized two satellite based NO_2 tropospheric column products, namely, OMI (Ozone Monitoring Instrument) (Levelt et al., 2006) and TROPOMI (TROPOspheric Monitoring Instrument) (Veefkind et al., 2012), to estimate NO_2 concentrations at ground level over Vietnam.

TROPOMI, launched in October 2017, is a satellite instrument on board the Copernicus Sentinel-5 Precursor satellite (S5P). It measures air quality, ozone, ultraviolet radiation, and aids in climate forecasts with high spatial resolution. TROPOMI provides daily and global coverage of multiple trace gases (such as NO_2, CO, SO_2, CH_4, CH_2O, O_3) and aerosol properties. Prior to Sentinel-5P, NASA's OMI on the Aura satellite had been observing the ozone layer and atmospheric pollutant gases, including NO_2, since October 2004. However, the daily OMI NO_2 product has a lower spatial resolution (13 × 24 km) compared to the more detailed NO_2 product from TROPOMI (3.5 × 5.5 km).

Both of OMI and TROPOMI data were obtained from the Multi-Decadal Nitrogen Dioxide and Derived Products from Satellites (MINDS) program (Lamsal et al., 2022a; Lamsal et al., 2022b). The goal of this project is to adapt OMI operating algorithms to other satellite devices, and to create and store consistent multi-satellite Level 2 and Level 3 NO_2 products. They adapt their well-validated OMI NO_2, cloud, and geometry-dependent surface reflectivity retrieval algorithms to satellite instruments that include SCIAMACHY, GOME-2, TROPOMI. The adaptation of OMI algorithms for these satellite data aims to provide consistent and long-term records suitable for analyzing global trends in NO_2. OMI MINDS NO_2 and TROPOMI MINDS NO_2 were both downloaded from NASA's open source (https://disc.gsfc.nasa.gov/). The data are listed in Supplementary Table S1.

2.3 Meteorological data

Meteorological parameters are the factors that have an important influence on the concentration of NO_2 pollutant over time. For example, high temperature can accelerate photochemical reactions thereby reducing NO_2 concentration; high relative humidity increases the conversion rate from NOx to secondary aerosols thereby also reducing NO_2 concentrations. In this study, we utilized meteorological maps generated by the Weather Research and Forecasting (WRF) model, which employed input data from the fifth generation of ECMWF reanalysis (ERA-5) obtained from (https://cds.climate.copernicus.eu) during 2019–2021. The spatial resolution of the input data was 0.25 ° × 0.25 ° with hourly temporal resolution. The meteorological data of the ERA-5 was used as the initial and boundary conditions for the simulation in the WRF model. The WRF configuration was set up with two nested domains over Vietnam, with spatial resolution of 15 and 5 km respectively. The output data of the model was meteorological maps (including Temperature, Humidity, WindSpeed, Planetary Boundary Layer Height - PBLH) with a frequency of 4 images/day at 0, 6, 12, 18 h (GMT+0) and a spatial resolution of 5 × 5 km. The data are listed in Supplementary Table S1.

2.4 Land use data

Land use factors are closely associated with the sources of emissions. For instance, regions characterized by high traffic density tend to exhibit elevated smog emissions from vehicles, leading to higher concentrations of NO_2. Conversely, areas covered with vegetation generally experience lower pollution levels compared to urbanized areas. In this study, we utilized the following data: normalized difference vegetation index (NDVI) map, road map. The data are listed in Supplementary Table S1.

The NDVI product used in this study is generated from Terra MODIS satellite images through the MOD13Q1 product, Collection 6, level 3, which has a spatial resolution of 250 m and a temporal resolution of 16 days (Didan, 2015). NDVI maps provide spatially and temporally consistent observations of vegetation status in the study area. In this study, we collected MOD13Q1 product during 2021 from NASA open source (https://search.earthdata.nasa.gov/search).

The road map used in this study was obtained from the latest OpenStreetMap (OSM) data in 2022, available in vector format and comprising road shapes. OSM is a community-driven mapping service that is freely accessible and open to the public. OSM widely employed in various applications within the geosciences, earth observation, and environmental sciences. OSM offers global map objects, including data types such as nodes (representing points on Earth), ways (polyline representations of road objects, buildings, *etc.*), relations (establishing relationships between objects), and tags (containing object-related information) (Vargas-Munoz et al., 2021).

3 Methods

This study developed daily NO_2 maps using a method shown in Figure 1. The input data included NO_2 data from monitoring stations, NO_2 tropospheric column density from satellites, meteorological maps from the WRF model, NDVI maps, and road maps. These data were preprocessed and integrated to

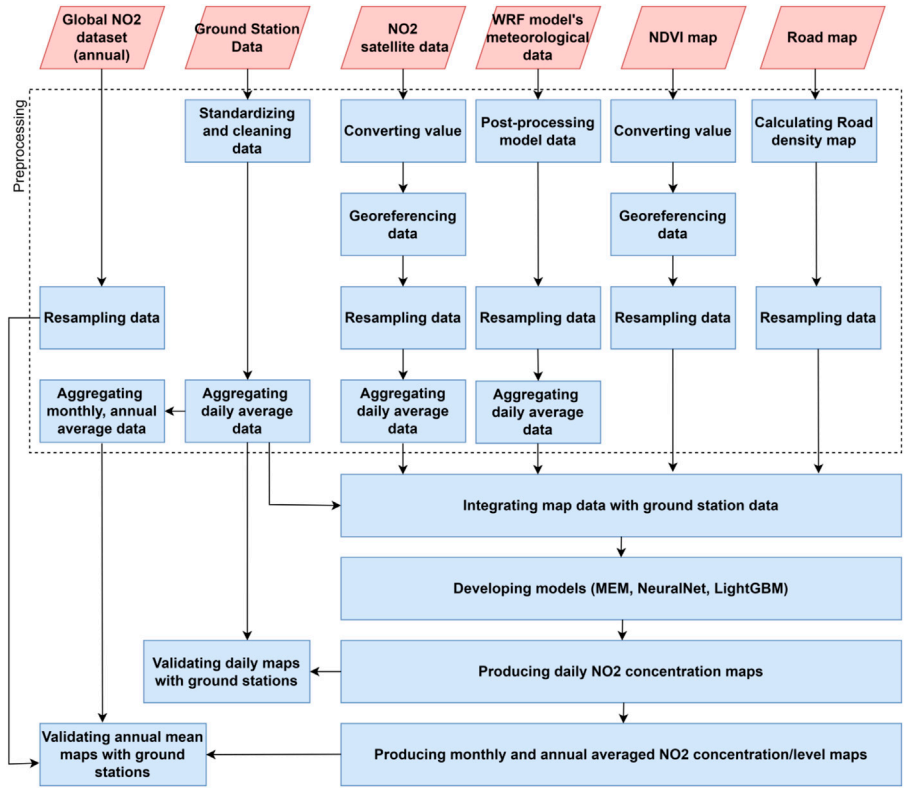

FIGURE 1
The methodological approach for estimating ground-level NO_2.

create a training dataset, which was used to develop statistical models for generating the daily NO_2 map. The daily NO_2 maps were then aggregated into monthly and annual averages, and validated using station observations and compared with the global NO_2 product.

3.1 Preprocessing data

The preprocessing of the monitoring station data, satellite images and ancillary data was similar to what we did for $PM_{2.5}$ pollutant published recently (Ngo et al., 2023). NO_2 concentration data from monitoring stations were standardized in uniform structure. After that, the data was cleaned and removed outliers. The process of removing outliers was carried out in the following steps: 1) Eliminating outliers by threshold. NO_2 observations with values exceeding 300 μg/m3 or less than 1 μg/m3 were discarded). 2) Using statistical methods to find outliers (too high/too low) compared to measured data in the neighboring period (±15 days). 3) Using the statistical method to find outliers (too high/too low) compared to the measured data in the neighboring period (±15 h), find out the outliers compared to the measured data measured at neighboring stations. 4) Finding outliers where the value does not change over a long period of time (Wu et al., 2018). These outliers were manually rechecked for accuracy. Subsequently, the hourly data were aggregated into daily, monthly, and annual averages for the purpose of data integration and modeling.

Multi-source satellite data, which are NO_2 tropospheric column density data from OMI, TROPOMI products and NDVI from the MOD13Q1 product, have different format, temporal and spatial resolutions. Preprocessing is required to convert satellite data into the same format and to project them in the same spatial grid. The preprocessing steps for the NO_2 and NDVI satellite images involve value extraction and transformation (converting value), geo-referencing, and resampling. Value extraction and transformation is the process of extracting related data layers and re-computing the values based on metadata information such as offset and scale factor of data. Geo-referencing means correlating the internal coordinate system of a map or an aerial image to a geographic coordinate system. In order to integrate multi-source data, a grid with uniform coverage and spatial resolution was defined. The grid covers the entire territory of Vietnam based on the WGS84 reference system and has cell size of 1 × 1 km. The satellite data were resampled and projected on this grid using the nearest resampling method for images with spatial resolution greater than 1 km (i.e., OMI, TROPOMI, meteorological maps) and the average resampling method for satellite images with resolution less than 1 km (i.e., MODIS NDVI, population density map). The GDAL tool was used to perform the above processes (GDAL, 2022). All the maps were then aggregated into daily maps for further calculation.

Quality flag bands were used to filter out low-quality pixels from the satellite products (OMI NO_2 and TROPOMI NO_2) to ensure the accuracy of the data. The bands used for filtering include "VcdQualityFlag" (even integer), "CloudFraction" (<0.3), and "qa_value" (>0.75), as recommended in previous studies (Lamsal et al., 2022a; Lamsal et al., 2022b). After quality control, the OMI and TROPOMI data were averaged on a daily basis to create a daily satellite combined dataset with a common grid (1 × 1 km grid).

The WRF model provides meteorological data in NetCDF format. The Unified Post Processing (UPP) Toolkit (NCEP UPP, 2022) was used to process the WRF model output data. UPP, which was developed at the National Center for Environmental Prediction (NCEP), has the capability of calculating various fields and interpolate them at different pressure levels from output data of the WRF model. We used the UPP tool to calculate temperature maps, humidity maps at 2 m height, planetary boundary layer height maps, wind speed at 10 m. Then, those data were resampled on the standard grid in order to be consistent with other satellite image products in the study area. These meteorological maps were then aggregated into daily mean maps for modeling.

NDVI, a MOD13Q1 product from Terra MODIS, was preprocessed similarly to those described for NO_2 maps, which were value extraction and transformation, geo-referencing, and resampling. The road map data was in vector format (shapefile), containing road lines and line characteristics. In order to use this feature as input of the model, the line density calculation was applied to convert the data into raster format (grids). It calculates a magnitude-per-unit area from polyline features, which fall within a radius around each cell (pixel). The radius is set approximately 1 km. The output image was then applied the nearest neighbor resampling method using the gdalwarp tool to get the same grid as the other maps.

3.2 Integrating data

Once the maps and station data were preprocessed, they were combined to create the training dataset. The aim was to establish the connection between the values on the maps and the observed NO_2 at the ground level. To ensure compliance with spatial and time constraints, the following measures were taken:

- Spatial constraint: The map data was extracted at the exact location of the ground station.
- Time constraint: The map data and ground-based NO_2 observations were synchronized by calculating the daily average values.

3.3 Modeling and validation

This study tested three different models: mixed effect model, neural network, and LightGBM. The MEM model has been widely used in the past to estimate pollution using satellite imagery and multi-source data. Recently, tree-based models have shown good results in estimating NO_2 maps. Therefore, in this study, we selected two machine learning models (MEM and LightGBM) to compare their performance. Additionally, we also wanted to experiment with a deep learning model. However, CNN-based models were not suitable for the current dataset, as complex deep learning models may not be suitable for sparse and limited data. Hence, we chose to experiment with a neural network model with multiple hidden layers and compared it with traditional machine learning models.

These models were fed with input parameters including NO_2 tropospheric column density (combined OMI and TROPOMI), meteorological data (humidity, PBLH), land cover (NDVI), and road density. Temperature and Wind Speed was not included in the input parameters due to its potential to create significant errors in estimating NO_2 concentrations in areas where ground monitoring stations are not installed in Vietnam. In other words, due to the uneven distribution of stations, the learned characteristics from the training dataset may not accurately reflect the patterns in areas without stations. For example, in mountainous regions with rocky terrain and dense forests (where there are no monitoring stations), the estimated pollution levels may appear higher than in flatland areas (with multiple monitoring stations, representing high emission areas).

The mixed effects model (MEM) is a type of land-use regression (LUR) model that consists of both fixed and random effect components. The formula for this model can be expressed as:

$$NO_{2\,i,j} = \sum_{k=1}^{N} \alpha_k X_{k,i,j} + (\alpha + \beta) \quad (1)$$

Where $NO_{2\,i,j}$ represents the estimated NO_2 concentration at spatial location j on day i. $X_{k,i,j}$ refers to the kth parameter at location j on day i, where N is the total number of parameters used in the model. The α_k, α coefficients denote the fixed effect component, which includes the slope and intercept of input parameters. The β coefficient represents the random effect of the intercept that varies from day to day.

LightGBM is a popular gradient boosting tree algorithm (Ke et al., 2017) used in machine learning. It utilizes a group of weak learners to improve the performance of the model. The regressor is optimized by adjusting hyper-parameters, such as the number of trees, the maximum tree depth, and learning rate, through the use of a grid search technique. The goal of this process is to improve the model's accuracy and reduce errors.

Neural network is a powerful method for modeling the complex and nonlinear relationships between inputs and outputs, which makes it suitable for studying atmospheric chemistry processes. It usually includes input, output, and hidden layers in its architecture (Nielsen, 2018). In this study, the neural network architecture was customized to fit the dataset size in terms of features and samples. During training and testing, the optimizer/learning rate, metric, and epochs were adjusted to optimize the performance of the model.

To assess the quality of the models, statistical indicators were used to compare the estimated NO_2 levels from the model with the actual NO_2 observations recorded at ground stations. The 10-fold CV method was employed to evaluate the performance of the model. After being trained and validated, the model was utilized to produce daily NO_2 concentration maps with a spatial resolution of 1 × 1 km. To evaluate its accuracy, the daily maps were compared with

TABLE 1 Models' evaluation results.

	Model	N	Pearson r	RMSE (μg/m^3)	MRE (%)
All data	Mixed Effect Model	9,027	0.66	9.39	54.01
	Neural Network	9,027	0.57	10.25	61.63
	LightGBM	9,027	0.87	6.28	34.65
10 Fold CV	Mixed Effect Model	903	0.56	10.46	59.29
	Neural Network	903	0.55	10.43	62.64
	LightGBM	903	0.77	7.93	42.6

ground station measurements using both temporal (daily mean) and spatial (pixel value extracted at station locations) constraints. In addition, to provide a more comprehensive analysis, daily ground measurements and NO_2 maps were aggregated into monthly and annual averages. The annual mean of our NO_2 maps was compared with the global NO_2 product (Anenberg et al., 2022) for the same study area, which provides annual average NO_2 datasets from 1990–2020 using a LUR method. The comparison involved evaluating the annual averages of our maps and the global product against ground station measurements of NO_2 taken in Vietnam from 2019–2021.

To compare and evaluate the models and maps, various statistical indicators were utilized, including the Pearson correlation coefficient (r), Root Mean Square Error (RMSE), and Mean Relative Error (MRE).

$$\text{Pearson r} = \frac{\sum_{t=1}^{n}(y_t - \bar{y})(x_t - \bar{x})}{\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - x_t)^2} \quad (3)$$

$$\text{MRE} = \frac{1}{N}\sum_{t=1}^{N}\frac{|y_t - x_t|}{y_t}.100\% \quad (4)$$

Here, x_t, y_t represent the estimated values from the model (or extracted from the map) and the measured values at the ground station, respectively. $\bar{x}$ and $\bar{y}$ are the respective average values of the two data series.

4 Results and discussion

4.1 Model validation

Supplementary Table S2 presents the selected parameters for each model. For the MEM model, the model structure has been presented in Section 3.3 and no parameters need to be adjusted. With the NN network, due to the small input dataset size (9,027 samples and 5 features), we designed a small size neural network consisting of 1 input layer, 3 hidden layers including 16 nodes, 32 nodes, 16 nodes, respectively. Adam optimizer was selected with the learning rate of 0.001. The metric used was mean squared error (MSE) and the epochs was set to 200. With the LightGBM model, through the grid search technique, we selected a set of parameters for the model which presented in the Supplementary Table S2.

Table 1 shows the evaluation results after setting up and training the models. Among the experimental models, the LightGBM model achieved the best performance, with a Pearson correlation coefficient of 0.87, RMSE of 6.28 μg/m^3, and MRE of 34.65%. In contrast, the MEM and Neural Network models had poorer quality. The LightGBM model also demonstrated superior performance in the 10-fold CV, with a Pearson correlation coefficient of 0.77, RMSE of 7.9 μg/m^3, and MRE of 42.6%. Based on these results, we selected the LightGBM model to estimate the daily NO_2 maps for Vietnam from 2019–2021, which were then aggregated into monthly and annual average maps.

4.2 Map validation

A comparison was made between the daily NO_2 maps and ground station measurements during the period of 2019–2021. The scatter plot depicted in Supplementary Figure S2 supports the findings presented in Table 1 regarding the model evaluation. The daily maps had a high correlation with the ground station observations, with Pearson r at 0.87, RMSE at 6.28 μg/m^3, MRE at 34.65% based on 9,027 samples. However, the evaluation results varied by stations as presented in Supplementary Table S3. Pearson r varied from 0.27 to 0.88 with lower values at stations in Vung Tau, Long An (SE and MRD region) and higher values in Bac Ninh, Quang Ninh, Ha Noi (RRD region). The RMSE varied from 2.1 to 10.1 μg/m^3. The stations with low RMSE values were located across regions, while stations with high RMSE were mostly located in Ha Noi, Bac Ninh, Quang Ninh (RRD). Furthermore, some stations located in the same province had highly different evaluation results, such as Bac Ninh, Hai Duong, Quang Ninh (RRD) and Gia Lai (Central Highland), indicating the need for further investigation.

Annual average NO_2 maps were created by aggregating daily NO_2 data from 2019 to 2021, as illustrated in Figure 2. The maps reveal that NO_2 was predominantly concentrated in the Red River Delta region in the North, along the North Central Coast, and in the Ho Chi Minh city area in the South. These regions are critical economic centers of Vietnam with high population density, heavy traffic, numerous industrial parks, and factories that generate significant NO_2 emissions. Across the country, the annual average NO_2 concentration varied from 4.4 to 36 μg/m^3 in 2019, 4.2 to 32.8 μg/m^3 in 2019 and 5.3 to 40.1 μg/m^3 in 2021. Notably, the national average concentration remained relatively stable between 2019 and 2021, indicating a persistent NO_2 pollution problem in Vietnam. Despite the implementation of social distancing measures

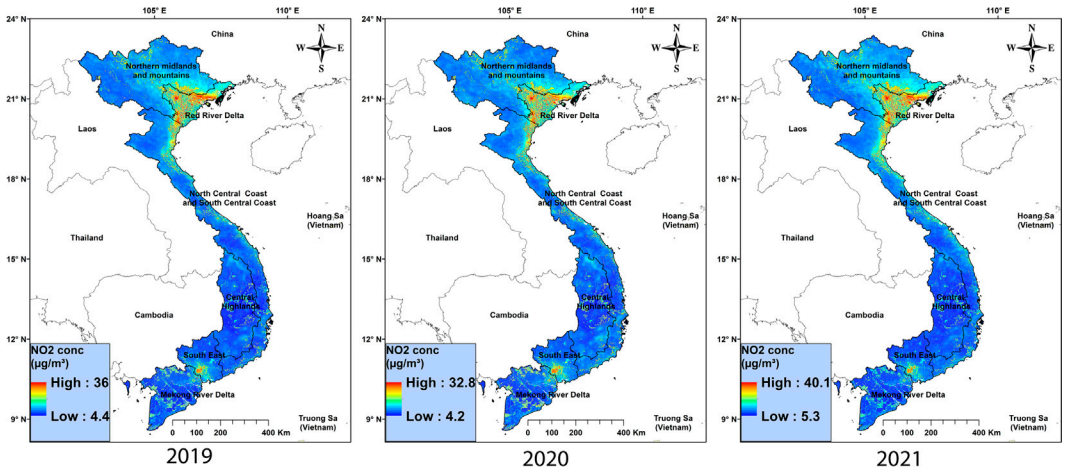

FIGURE 2
Annual mean ground-level NO_2 maps from 2019 to 2021.

TABLE 2 Comparison of validation results for ours and the global annual mean maps to ground station values.

Time	Study	N	Pearson *r*	RMSE (μg/m³)	MRE (%)
2019–2020	(Anenberg et al., 2022)	33	0.27	13.3	57.4
2019–2020	This study	33	0.95	2.1	8.6
2019–2021	This study	81	0.95	2.27	9.79

in response to the COVID-19 pandemic in 2019 and 2020 in Vietnam, there was not a significant variation in the annual mean NO_2 levels measured at stations. This lack of variation resulted in no significant changes in the annual NO_2 maps over the years (see Table 2; Supplementary Figure S3).

In Supplementary Figure S3, a detailed comparison is presented between the annual average NO_2 concentration maps for the years 2019, 2020, and 2021, and the ground stations located in Vietnam. It is noteworthy that the number of stations used for annual map assessment is less than that used for daily map assessment. This is because, when aggregating daily data into an annual average, any station that did not have more than 50% of the data for the year was discarded and not used for evaluation. Furthermore, in 2019, only three stations were evaluated, whereas this number increased to 30 in 2020 and to 48 in 2021. The difference between the annual maps and the ground stations varied from −2.5 μg/m³ (Quang Ninh - RRD) to 0.6 μg/m³ (Ha Noi - RRD) in 2019; −3.96 μg/m³ (Bac Ninh - RRD) to 4.6 μg/m³ (Ha Noi - RRD) in 2020; −4.12 μg/m³ (Ha Noi–RRD) to 8.13 μg/m³ (Bac Ninh–RRD) in 2021.

To ensure a thorough assessment, we compared the quality of our annual maps from 2019 to 2021, not only against ground stations, but also against the annual global product (2019–2020) developed by Anenberg et al. (2022). Table 2 displays the findings. Our annual maps showed markedly superior quality in comparison to both the global annual maps and the ground stations. Specifically, we achieved a Pearson correlation coefficient of 0.95, an RMSE of 2.1 μg/m³, and an MRE of 8.6%, while the global annual maps achieved only a Pearson r of 0.27, an RMSE of 13.3 μg/m³, and an MRE of 57.4%. Additionally, our map from 2019 to 2021 had a Pearson *r* of 0.95, an RMSE of 2.27 μg/m³, an MRE of 9.79%, and 81 samples, indicating the high quality of the annual NO_2 maps in this study and the potential of this approach to develop NO_2 maps from multi-satellite images over Vietnam.

5 Conclusion

In this study, daily NO_2 maps at 1 × 1 km over Vietnam were created using OMI and TROPOMI satellite images as well as auxiliary data from 2019–2021. Three models were experimented, including MEM, NN, and LightGBM, with LightGBM proving to have the best quality (Pearson *r* of 0.87, RMSE of 6.28 μg/m³, MRE of 34.65%). The LightGBM model was used to generate the daily NO_2 maps, which were validated against ground stations and found to be accurate. However, the quality of the maps varied by station, with Pearson r ranging from 0.27 to 9.88 and RMSE ranging from 2.1 to 10.1 μg/m³ between 2019–2021. The daily maps were then combined to produce monthly and yearly average maps. Our annual average map was compared to a global product and ground stations, and it was found to have superior quality with Pearson r of 0.95, RMSE of 2.27 μg/m³, MRE of 9.79%, and 81 samples. This is the first study on constructing NO_2 concentration maps in Vietnam using multi-source satellite data. The study encountered challenges such as uneven distribution of monitoring stations in the research area and limitations posed by cloud coverage on NO_2 satellite data (OMI, TROPOMI). Further exploration of these issues is needed in future research to enhance the quality of the maps.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

TXN: methodology, software, validation, formal analysis, writing—original draft, and visualization. HP: writing—review and editing. TTNN: conceptualization, methodology, validation, writing—original draft, and supervision. All authors contributed to the article and approved the submitted version.

Funding

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2019.331.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2023.1187592/full#supplementary-material

References

Anenberg, S. C., Mohegh, M., Goldberg, L., Kerr, H. K., Brauer, B., Burkart, K., et al. (2022). Long-term trends in urban NO2 concentrations and associated paediatric asthma incidence: Estimates from global datasets. *Lancet Planet. Health* 6 (1), e49–e58. doi:10.1016/S2542-5196(21)00255-2

Didan, K. (2015). "MOD13Q1 MODIS/Terra vegetation indices 16-day L3 global 250m SIN grid V006," in *Distributed by NASA EOSDIS land processes DAAC* (United States: United States Geological Survey). doi:10.5067/MODIS/MOD13Q1.006

GDAL (2022). GDAL documentation 2022. Available at: https://gdal.org/programs/gdalwarp.html.

Grzybowski, P. T., Markowicz, M., and Musiał, J. P. (2023). Estimations of the ground-level NO2 concentrations based on the sentinel-5P NO2 tropospheric column number density product. *Remote Sens.* 15 (2), 378. doi:10.3390/rs15020378

Kang, Y., Choi, H., Im, I., Park, S., Shin, M., Song, C-K., et al. (2021). Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia. *Environ. Pollut.* 288, 117711. doi:10.1016/j.envpol.2021.117711

Ke, G., Qi, M., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree." in *Advances in neural information processing systems*. Long Beach, CA: Curran Associates, Inc.

Lamsal, L. N., Krotkov, N. A., Marchenko, S. V., Joiner, J., Oman, L., Alexander, V., et al. (2022a). *OMI/Aura NO2 tropospheric, stratospheric and total columns MINDS 1-orbit L2 swath 13 Km x 24 km*. Greenbelt, Maryland: Goddard Earth Sciences Data and Information Services Center GES DISC. NASA Goddard Space Flight Center.

Lamsal, L. N., Krotkov, N. A., and Marchenko, S. V. (2022b). *TROPOMI/S5P NO2 tropospheric, stratospheric and total columns MINDS 1-orbit L2 swath 5.5 Km x 3.5 km*. Greenbelt, Maryland: Goddard Earth Sciences Data and Information Services Center. NASA Goddard Space Flight Center.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., et al. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environ. Sci. Technol.* 51 (12), 6957–6964. doi:10.1021/acs.est.7b01148

Lee, H. J., and Koutrakis, P. (2014). Daily ambient NO2Concentration predictions using satellite ozone monitoring instrument NO2Data and land use regression. *Environ. Sci. Technol.* 48 (4), 140204134232009–140204134232011. doi:10.1021/es404845f

Levelt, P. F., van den OordVan Den Oord, H. J., Dobber, M. R., Malkki, A., Huib Visser, H., Johan de Vries, J., et al. (2006). The ozone monitoring instrument. *IEEE Trans. Geosci. Remote Sens.* 44 (5), 1093–1101. doi:10.1109/TGRS.2006.872333

Long, S., Wei, X., Zhang, F., Zhang, R., Xu, J., Wu, K., et al. (2022). Estimating daily ground-level NO2 concentrations over China based on TROPOMI observations and machine learning approach. *Atmos. Environ.* 289, 119310. doi:10.1016/j.atmosenv.2022.119310

NCEP UPP (2022). NCEP unified Post processing system (UPP). Available at: https://dtcenter.org/community-code/unified-post-processor-upp.

Ngo, T. X., Pham, H. V., Phan, H. D. T., Nguyen, A. T. N., To, H. T., and Nguyen, T. T. N. (2023). A daily and complete PM2.5 dataset derived from Space observations for Vietnam from 2012 to 2020. *Sci. Total Environ.* 857, 159537. doi:10.1016/j.scitotenv.2022.159537

Nguyen, T., Bui, H. Q., Pham, H. V., Luu, H. V., Man, C. D., Pham, H. N., et al. (2015). Particulate matter concentration mapping from MODIS satellite data: A Vietnamese case study. *Environ. Res. Lett.* 10 (9), 095016. doi:10.1088/1748-9326/10/9/095016

Nielsen, M. A. (2018). *Neural networks and deep learning*. Oxford: Determination Press.

Paraschiv, S., Constantin, D. E., Paraschiv, S. L., and Constantin, M. (2017). OMI and ground-based *in-situ* tropospheric nitrogen dioxide observations over several important European cities during 2005–2014. *Int. J. Environ. Res. Public Health* 14 (11). doi:10.3390/ijerph14111415

Vargas-Munoz, J. E., Srivastava, S., Tuia, D., and Falcao, A. X. (2021). OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience Remote Sens. Mag.* 9 (1), 184–199. doi:10.1109/MGRS.2020.2994107

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., et al. (2012). 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.* 120, 70–83. doi:10.1016/j.rse.2011.09.027

Wang, W., Liu, X., Bi, J., and Liu, Y. (2022). A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology. *Environ. Int.* 158, 106917. doi:10.1016/j.envint.2021.106917

Wu, H., Tang, X., Wang, Z., Wu, L., Lu, M., Wei, L., et al. (2018). Probabilistic automatic outlier detection for surface air quality measurements from the China national environmental monitoring network. *Adv. Atmos. Sci.* 35 (12), 1522–1532. doi:10.1007/s00376-018-8067-9