Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrating Sentinel-5P Satellite Observations and Machine Learning for Estimating Ground-Level NO 2 Concentrations: A Case Study of Urmia, Iran

Samin Rahimzadeh
Urmia University
Amir Asadi Vaighan
a.asadi@urmia.ac.ir

Urmia University
Amin Rahimzadeh
Islamic Azad University of Urmia

Article

Keywords: Nitrogen dioxide (NO2), Sentinel-5P, Machine learning, Gradient Boosting, Air quality monitoring, Urmia

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11201562/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrating Sentinel-5P Satellite Observations and Machine Learning for Estimating Ground-Level NO_2 Concentrations: A Case Study of Urmia, Iran

Samin Rahimzadeh[1], Amir Asadi.Vaighan[1]*, Amin Rahimzadeh[2]

[1]Department of Civil Engineering, Faculty of Engineering, Urmia University, Urmia, Iran.
st_s.rahimzadeh@urmia.ac.ir

Department of Architecture, Faculty of Engineering, Islamic Azad University, Urmia Branch, Urmia, Iran, Email: amin.rahimzadeh@iau.ir

Correspondence to: Amir Asadi Vaighan (a.asadi@urmia.ac.ir)

Abstract:

The process of monitoring nitrogen dioxide (NO_2) levels in areas that lack sufficient ground-based measurement stations presents significant difficulties for effective air quality control. The TROPOMI sensor on Sentinel-5P satellite provides a comprehensive spatial monitoring system but requires ground-based data and meteorological information to correctly analyze tropospheric column measurements and surface-level concentration data. The researchers assessed how effectively satellite data and ground-based information and meteorological data work together within a machine learning system to predict surface NO_2 levels in Urmia city from 2019 to 2024. The researchers used Google Earth Engine to extract Sentinel-5P satellite products which included NRTI and OFFL and they combined this information with data from the Urmia Air Quality Monitoring Station and the Urmia Airport Synoptic Station. The machine learning models which included Decision Tree Random Forest Extra Trees Gradient Boosting XGBoost CatBoost and AdaBoost were trained using an 80% training and 20% testing split after the researchers completed data normalization through incomplete and outlier record removal and Pearson correlation coefficient ($|R| \geq 0.10$) variable selection. The researchers used R^2, RMSE, and MAE metrics to evaluate performance. The tested models showed Gradient Boosting as the most effective model because it produced R^2=0.9961, MAE=0.0212 and RMSE=0.0578 results for NRTI data and R^2=0.9959, MAE=0.0290 and RMSE=0.0634 results for OFFL data. The researchers used scatter plots and residuals and error distributions to establish that their study results had no systematic bias and no overfitting problems. The results demonstrate that Sentinel-5P data combined with ground-based measurements and meteorological data through Boosting algorithms, especially Gradient Boosting, enables accurate surface-level NO_2 estimation in cities that lack ground monitoring stations. The development of region-specific models requires local data in Urmia because the city has distinct climatic conditions and pollutant emission patterns.

Keywords: Nitrogen dioxide (NO_2), Sentinel-5P, Machine learning, Gradient Boosting, Air quality monitoring, Urmia.

1. Introduction

The current century considers air pollution as its most serious environmental problem which affects human health and ecosystems and sustainable urban development. The industrial growth and rapid urbanization and transportation fleet expansion drive increasing air pollution when they operate in multiple cities worldwide (Khomenko et al., 2021). Nitrogen dioxide (NO_2) functions as a major atmospheric pollutant which results from fossil fuel combustion and urban traffic while it contributes to the development of tropospheric ozone and secondary particulate matter (Ghozikali et al., 2016).

Epidemiological studies have shown that long-term exposure to pollutants such as NO_2 is associated with increased premature mortality and the incidence of respiratory and cardiovascular diseases (Vineis et al., 2020; Khomenko et al., 2021). Urban assessments show that respiratory diseases burden a large portion of the population because of NO_2 and SO_2 pollutant concentrations (Ghozikali et al., 2016). Severe air pollution episodes throughout history have caused fatalities which reached extreme levels during the Great London Smog of 1952 when 4,000 people died within the first few days and over 12,000 people died during the following months which demonstrates the necessity to monitor and control the pollutants released during fossil fuel combustion (Bell & Davis, 2001).

The network of air quality monitoring stations in various regions lacks operational efficiency because their stations do not provide complete spatial coverage and their collected data suffers from both time duration breaks and accessibility problems. Cities with fewer monitoring stations have this problem because they cannot evaluate pollution levels and execute urgent response measures. Satellite remote sensing functions as an effective solution to monitor atmospheric pollution across large areas according to research conducted by Dubovik and his colleagues in 2021. The Aura/OMI satellite and Sentinel-5P satellite which includes the TROPOMI sensor, enable researchers to detect tropospheric NO_2 levels across extensive areas while obtaining frequent data updates according to Krotkov and Wieczorek in 2016 and 2023.

The satellite observations do not show actual surface concentration levels which creates a need for researchers to combine satellite data with ground measurements and meteorological data for accurate NO_2 surface level estimation (Kim et al., 2021). The development of machine learning techniques allows researchers to discover nonlinear connections between multiple datasets while they use these methods to achieve better accuracy in estimating pollutants (Siddique et al., 2024).

Recent studies show that satellite and ground-based data integration using machine learning models leads to accurate surface NO_2 estimation. The Random Forest model which used multiple data sources in China achieved R^2 value of 0.64 at national level according to Dou et al. (2020). The satellite observation frameworks which use machine learning showed R^2 results between 0.6 and 0.8 together with RMSE measurements of several micrograms per cubic meter, which outperformed standard techniques. The studies Chan et al. (2021), Long et al. (2022), and Chi et al. (2022) confirmed that TROPOMI observations combined with machine learning algorithms

successfully model the spatial and temporal changes of NO_2. The statistical and machine learning methods used in the United Arab Emirates case study proved that machine learning models outperformed traditional approaches for NO_2 temporal variation reconstruction according to Al Yammahi and Aung (2023). The study results showed that artificial intelligence models including Random Forest produced their lowest error rates through an RMSE of 6.01 and an MAE of 3.71. The research found that the top models achieved MAPE values close to 10%, which demonstrated significant progress over traditional statistical methods for air quality assessment in dry desert environments.

Multiple current investigations have utilized satellite data through Google Earth Engine (GEE) platform which enables scientists to obtain TROPOMI and OMI sensor data at high speed and reliable intervals. The machine learning method which combines satellite data with ground measurements and meteorological data has produced acceptable model evaluation metrics across various cities and different weather conditions and multiple timeframes as shown in Table (1). The differences among emission sources and traffic patterns and geographic features and weather conditions between different areas make it unsafe to apply research findings from one location to another which makes it necessary to analyze each area separately through independent research. The combination of TROPOMI data with meteorological data enables researchers to produce detailed maps of surface NO_2 pollution which shows how different pollution levels affect urban areas according to multiple studies conducted at the regional level (Kim et al 2021; Long et al 2022; Chi et al 2022).

Most existing studies have studied national and metropolitan areas which cover East Asia Europe and North America while researchers have conducted limited multi-source studies about medium-sized developing cities especially those in Iran. The assessment of satellite observation accuracy to actual surface conditions in those areas encounters problems because ground monitoring data exists only during certain periods. The existing research requires further regional investigations which need to study how Sentinel-5P satellite data correlates with ground measurements during various climatic and topographical conditions.

The present study investigates how Sentinel-5P satellite data from Google Earth Engine platform relates to surface NO_2 values while examining machine learning algorithm's ability to predict pollutant changes in Urmia from 2019 to 2024. The main research question investigates whether satellite data and limited ground observations together with meteorological variables can accurately estimate NO_2 distribution patterns across Urmia which has specific climatic and emission characteristics. The study conceptual framework receives additional support from previous research findings about satellite and ground-based data integration methods used to analyze pollutants (Ghanbarzadeh Lak et al., 2025).

Table 1. Overview of previous research integrating satellite remote sensing and meteorological variables for air pollution mapping via machine learning.

Source	Satellite & Sensor	Study Area & Period	Pollutant(s)	Multi-source Input Data	Selected Machine Learning Method(s)	Performance Metrics / Evaluation Results
Chan et al. (2021)	S5P[1] / TROPOMI	Germany (2018–2019)	NO_2	Satellite data, ground observations, traffic, meteorology, elevation	XGBoost	$R^2 = 0.79$ RMSE = 4.7
Chi et al. (2022)	Aura / OMI	China (2013–2019)	NO_2	Satellite data, boundary layer height (BLH), humidity, temperature	Spatiotemporal models & Random Forest	$R^2 = 0.62$ - 0.78
Ahmad et al., 2024	GEMS	China (2019–2021)	NO_2	GEMS satellite data, meteorology (temp, humidity, wind), land-use info, nested ML	Nested XGBoost	$R^2 = 0.79$, RMSE = 5.86 µg/m³
Long et al. (2022)	S5P / TROPOMI	China (2019–2020)	NO_2	NO_2Satellite data, 2m temperature, planetary boundary layer height (PBLH), pressure, evapotranspiration, humidity	Random Forest	$R^2 = 0.77$ RMSE = 7.88
Al Yammahi (2023)	Ground Stations	UAE (2018–2021)	NO_2	Historical concentration profiles, meteorological parameters	Random Forest, XGBoost, LSTM	RMSE = 6.01 MAE = 3.71
Kim et al. (2021)	S5P / TROPOMI	Europe (2018)	NO_2	NO_2Air temperature, wind components (u, v), pressure, humidity, cloud cover	Gradient Boosting	$R^2 = 0.81$
Siddique et al. (2024)	OMI & TROPOMI	Review Study (Global scale)	NO_2	NO_2 Hybrid (Meteorology, traffic, spatiotemporal features)	RF, DL, Ensemble Learning	$R^2 = 0.70$ - 0.85
Mauro et al. (2023)	S5P / TROPOMI	Italy (Campania) (2019–2020)	NO_2	NO_2 Satellite-derived data, meteorology (temp, humidity, wind), station geospatial data	CatBoost	RMSE = 0.0242 (on 43 stations)
Ghanbarzadeh Lak et al. (2025)	S5P / TROPOMI	Iran (Urmia, West Azerbaijan) (2019–2024)	SO_2	SO_2Air temperature, relative humidity, evapotranspiration, dew point, cloud cover, station air pressure	Decision Tree Random Forest Extra Trees Gradient Boosting XGBoost CatBoost AdaBoost	$R^2 = 0.89$ RMSE = 5.7
Zhan et al. (2018)	Aura / OMI	China (2012 – 2016)	$PM_{2.5}$	Air temperature, wind speed, humidity, pressure, solar radiation	Gradient Boosting	$R^2 = 0.62$ - 0.75

[1] Sentinel-5P

2. Methodology

2-1. Study Area

The city of Urmia, which serves as the capital of West Azerbaijan Province, extends from 37°06′ N latitude to 37°48′ N latitude and from 45°12′ E longitude to 45°48′ E longitude (Figure 1). The city lies adjacent to Lake Urmia, which is the largest inland lake in Iran and the second-largest saline lake in the Middle East. The climate of Urmia experiences Atlantic Ocean and Mediterranean Sea air masses, which create semi-arid weather conditions. The region receives about 350 millimeters of rainfall annually (Ahmady-Birgani et al., 2022; Nouri et al., 2021).

2-2. Data Collection

The Urmia AQ[2] Monitoring Station (Station No. 3 – Figure 1, (A)) provided ground-level NO_2 concentration data to researchers (supplementary file: gap.xlsx). The station is located at 37.5544° N latitude and 45.0484° E longitude, and daily measurements were recorded over a five-year period from 2019 to 2024. The Urmia Synoptic Airport Station (Urmia Meteorological Station) provided meteorological data (Figure 1, (B)), which included temperature measurements at various levels, wet-bulb temperature, dew-point temperature, relative humidity, actual and saturation vapor pressure, reduced pressure to sea level, station-level pressure, wind speed and direction, evaporation from pan, daily sunshine hours, 6- and 24-hour precipitation, cloud cover, solar radiation, and horizontal visibility (supplementary file : heatmap.xlsx). The data was obtained through official channels from the Iran Meteorological Organization website.

The TROPOMI sensor on the Sentinel-5P satellite provided satellite-derived NO_2 data, which researchers obtained for Station No. 3 coordinates through the Google Earth Engine (GEE) platform ecosystem (supplementary files: NTRI.xlsx and OFFL.xlsx). The satellite data validated during the five-year period from 2019 to 2024 and served as the basis for subsequent analyses. (We pulled these data from Google Earth Engine.

2-2. Data Preprocessing

The process started with the detection of missing data and time intervals which needed to be addressed before conducting analysis on the complete dataset. The team identified outliers which had the potential to decrease machine learning model performance and proceeded to eliminate them from the dataset. The Z-Score method was used to identify outliers because their values stem from instrument faults or exceptional circumstances or measurement errors.

The researchers calculated daily average values for the data to create a unified dataset which combined satellite and ground-based data. The researchers used Standard Scaler method to perform data normalization which established uniform measurement standards for all variables in

2 Air Quality

the study. The researchers merged two datasets through their time index after they completed the process of matching data to establish a unified dataset for machine learning algorithms.

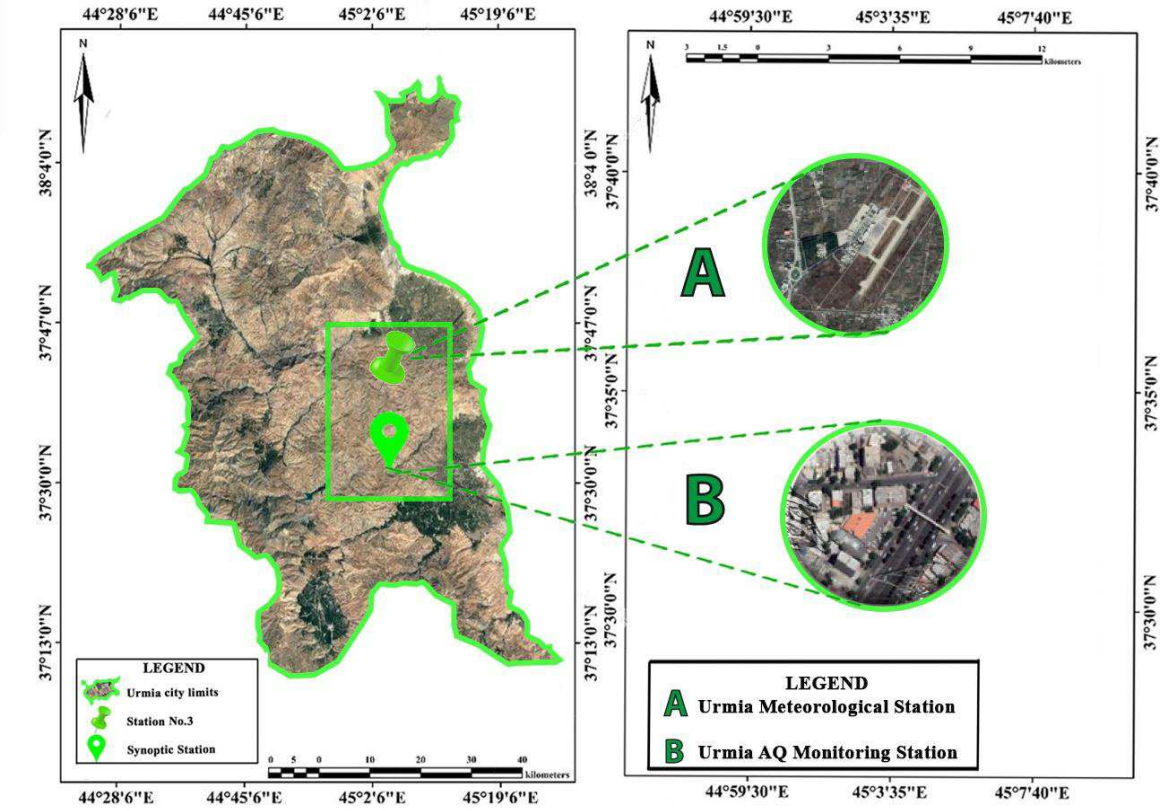


Figure 1. Study Area: Urmia city limits; (A) Urmia AQ Monitoring Station (Station No.3); (B) Urmia Meteorological Station (Synoptic Station).

2-3. Machine Learning

The researchers used machine learning which represents an advanced field of artificial intelligence to study the combined ground-based and satellite NO_2 concentration data. The datasets were aggregated as daily averages and records lacking complete daily information were excluded to enhance model performance and result accuracy (Budach et al., 2022). The researchers used Python version 3.8 through Jupyter Notebook to run the machine learning algorithms. Jupyter Notebook provides an effective platform for data processing and model development because it allows users to run code blocks separately while observing the results at the same time (Kazemi Garajeh et al., 2023).

The research team used the essential Python libraries NumPy and Pandas and Matplotlib and Scikit-learn to handle and process their data. NumPy enabled researchers to prepare machine

learning data because it provides advanced numerical computation tools and multidimensional array management functions. The data preprocessing process became more efficient because Pandas provides Data Frames which enable reading and cleaning and manipulating data. The researchers used Matplotlib to create visual representations of data which included various 2D and 3D plots that helped them understand how different variables related to each other.

The system uses Scikit-learn to execute its algorithms which provide complete access to their classification and regression and clustering methods. The Scikit-learn library provides complete documentation together with streamlined computing processes which help users achieve both model performance and operational efficiency. The system used these tools to create a unified system which successfully built machine learning models and tested their performance while delivering trustworthy and accurate outcomes.

Researchers began their modeling work with tree-based algorithms which included Decision Tree and Random Forest and Extra Trees because these methods excel at handling complex datasets and forecasting pollutant emissions. The researchers used multiple boosting-based algorithms which included Gradient Boosting and XGBoost and CatBoost and AdaBoost for their analysis. The methods create strong models through their combination of multiple weak models which results in precise predictions for NO_2 concentrations in complex data sets.

The Pearson correlation coefficient (R) was used to measure air pollution levels. The statistical index common in research studies assesses how two continuous variables share a direct relationship. The Pearson correlation coefficient ranges from −1 to +1, where +1 indicates a perfect positive correlation which −1 represents a perfect negative correlation and zero denotes the absence of a meaningful linear relationship between variables. Positive values of R indicate that an increase in one variable is associated with an increase in the other, whereas negative values suggest that an increase in one variable corresponds to a decrease in the other (Tavella et al., 2025). The study applied Pearson correlation coefficient to evaluate how NO_2 levels connected with meteorological data and other air pollutants which were measured at ground monitoring stations through Python programming (Equation 1).

To evaluate the accuracy and performance of the machine learning models applied in this research, the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were adopted as the primary performance assessment metrics. The coefficient of determination shows how well a model can predict the results of a dependent variable based on its independent variables, with its output values extending from 0 to 1. The model outputs show better correlation with actual data when their values approach 1, which demonstrates that the model can explain most of the changes in nitrogen dioxide (NO_2) levels, while lower values indicate the model performs worse (Equation 2). The RMSE value shows the model error size because it calculates the difference between estimated and actual results, where the model achieves higher accuracy through small RMSE values but shows greater modeling error through large value differences (Equation 3). The MAE method calculates total error through its MAE process, which

calculates mean absolute differences between predicted outcomes and actual data yet shows reduced effects from outliers (Equation 3). The machine learning models achieved satisfactory performance in analyzing NO_2-related data because they combined low RMSE and MAE values with high R^2 values (Soleimany et al., 2021).

$$R = \frac{\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^{n} (b_i - \bar{b})^2}} \quad [1]$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad [2]$$

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \quad [3]$$

$$MAE = \frac{\sum_i |y_i - \hat{y}_i|}{n} \quad [4]$$

The paired values of the two variables for the i-th observation used to calculate the Pearson correlation coefficient are represented by the elements (a_i) and (b_i) while the arithmetic mean values of variables (a) and (b) are denoted by the symbols ($\bar{a}$) and ($\bar{b}$).The observed value for the i-th sample is shown as (y_i) while the machine learning model predicts the corresponding value through ($\hat{y}_i$)and ($\bar{y}$) represents the average of all observed values in the study which used (n) samples for its research work.

3. Results and Discussion

3.1. Data Quality Assessment and Feature Selection

Figure 2 shows a heatmap which displays the ground-level NO_2 data that was not collected during the time period from 2019 to 2024. The horizontal axis represents pollutant concentrations, while the vertical axis indicates the sampling years. The color scale on the right of the figure shows the percentage of missing data, where lighter colors correspond to higher proportions of missing values and darker colors indicate the presence of complete data.

The heatmap analysis shows that ground-level NO_2 data collection is incomplete because there are multiple days without any measurement data. The years 2021 and 2022 show significant data deficiencies because there are complete months without any trustworthy information. The year 2024 shows no ground-level NO_2 data because monitoring stations either stopped working or suspended their operations for the entire year. The existing situation demonstrates how air quality monitoring stations face operational challenges because they lack enough resources to provide

continuous monitoring of air quality, which hinders research efforts and air quality forecasting activities.

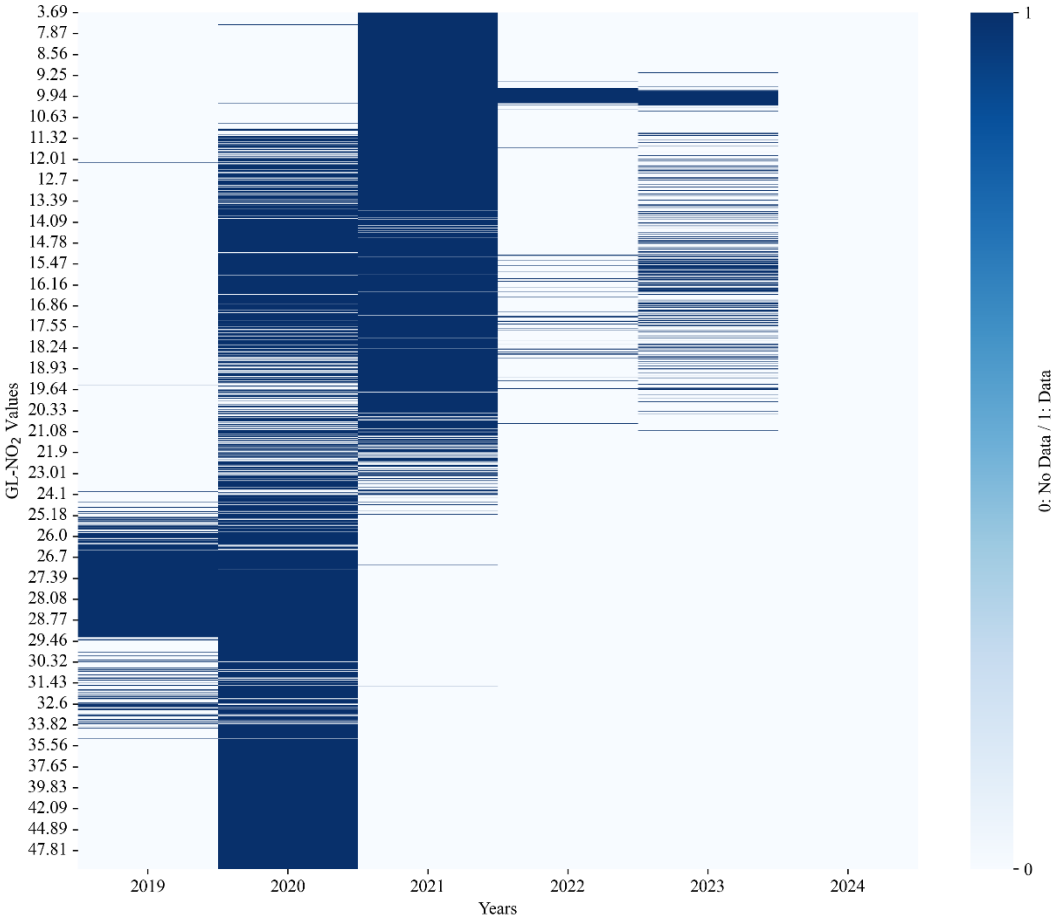


Fig 2- Heatmap of Missing Ground-Based Data for Air Pollutants During the Study Period

The satellite data stands as essential data because ground-level NO_2 measurements face scarcity and monitoring stations have operational restrictions. Satellite observations can capture the spatial and temporal variations of NO_2 concentrations and enable monitoring in areas lacking ground stations. The combination of these data with meteorological data together with machine learning algorithms creates a system that enables more thorough and precise time-series assessments while delivering an effective solution for air pollution control and environmental decision-making in urban environments.

The researchers conducted Pearson correlation analysis to find factors which improve the relationship between ground-based and satellite-derived NO_2 concentrations. The heatmap displayed meteorological data together with ground-level pollutant measurements which the researchers used for their analysis (Figure 3). The Urmia Airport synoptic station provided

meteorological data which included horizontal visibility (vv) and cloud cover (n) and wind direction (dd) and wind speed (ff) and air temperature (t) and dew point temperature (td) and sea-level pressure (p) and station pressure (p_0) and 6-hour and 24-hour precipitation (rrr and rrr24) and relative humidity (u) and minimum and maximum air temperature (tmin and tmax) and daily global radiation (radglo24) and sunshine duration (sshn) and evapotranspiration (evt) and wet-bulb temperature (twet) and maximum wind speed and direction (ffmax and ddmax) and vapor pressure (ew) and saturated vapor pressure (ews). The researchers included multiple ground-level pollutant concentrations in their assessment. The researchers used nitrogen oxides (NO_X) and nitric oxide (NO) and sulfur dioxide (SO_2) and particulate matter with diameters less than 10 μm (PM_{10}) and particulate matter with diameters less than 2.5 μm ($PM_{2.5}$) and carbon monoxide (CO) to study how these pollutants interact with ground-level NO_2 concentrations. The system allowed researchers to study how different pollutants affected the strength of the connection between ground observations and satellite NO_2 data.

The heatmap results demonstrated that nitrogen oxides (GL- NO_X) showed the strongest link to ground-level $NO_2$with a correlation coefficient of +0.96 and nitric oxide (GL-NO) with +0.90, which demonstrated an extremely high correlation because both pollutants came from the same sources. Sulfur dioxide (GL- SO_2) showed a moderate correlation with ground-level NO_2 with a correlation coefficient of 0.42. The meteorological parameters demonstrated their strongest links to ground-level NO_2 through their associations with relative humidity (u) (+0.22), evapotranspiration (evt) (−0.20), cloud cover (n) (+0.15), air pressure (p) (+0.14), wind speed (ff) (−0.14), air temperature (t) (−0.13), sunshine duration (sshn) (−0.12), saturated vapor pressure (ews) (−0.15), and 6-hour precipitation (rrr) (+0.10).

The study used Pearson correlation coefficient absolute values to determine which parameters would be considered essential, while all variables in the study with correlation coefficients between 0.10 and 1.0 were treated as effective measurement tools. The variables GL- NO_X, GL-NO, GL-SO_2, n, ff, t, p, rrr, u, sshn, evt, and ews were selected for subsequent analyses.

The analysis did not continue with GL-CO because it showed enough correlation to be valid but critical data points were missing which created uncertainty about the results. The researchers removed the variables rrr24 and tmax from their study after the selected range showed them to have high correlation with rrr and t. The model building process required researchers to remove p0 because its strong correlation with p created multicollinearity issues.

The final parameter set emerged from a data-driven approach which used the correlation matrix results to identify variables that had independent effects on subsequent analyses and machine learning models. The relationship between ground-level NO_2 variability and Sentinel-5P satellite observations can be strengthened by including selected pollutant concentrations, especially NO_X and NO together with meteorological parameters.

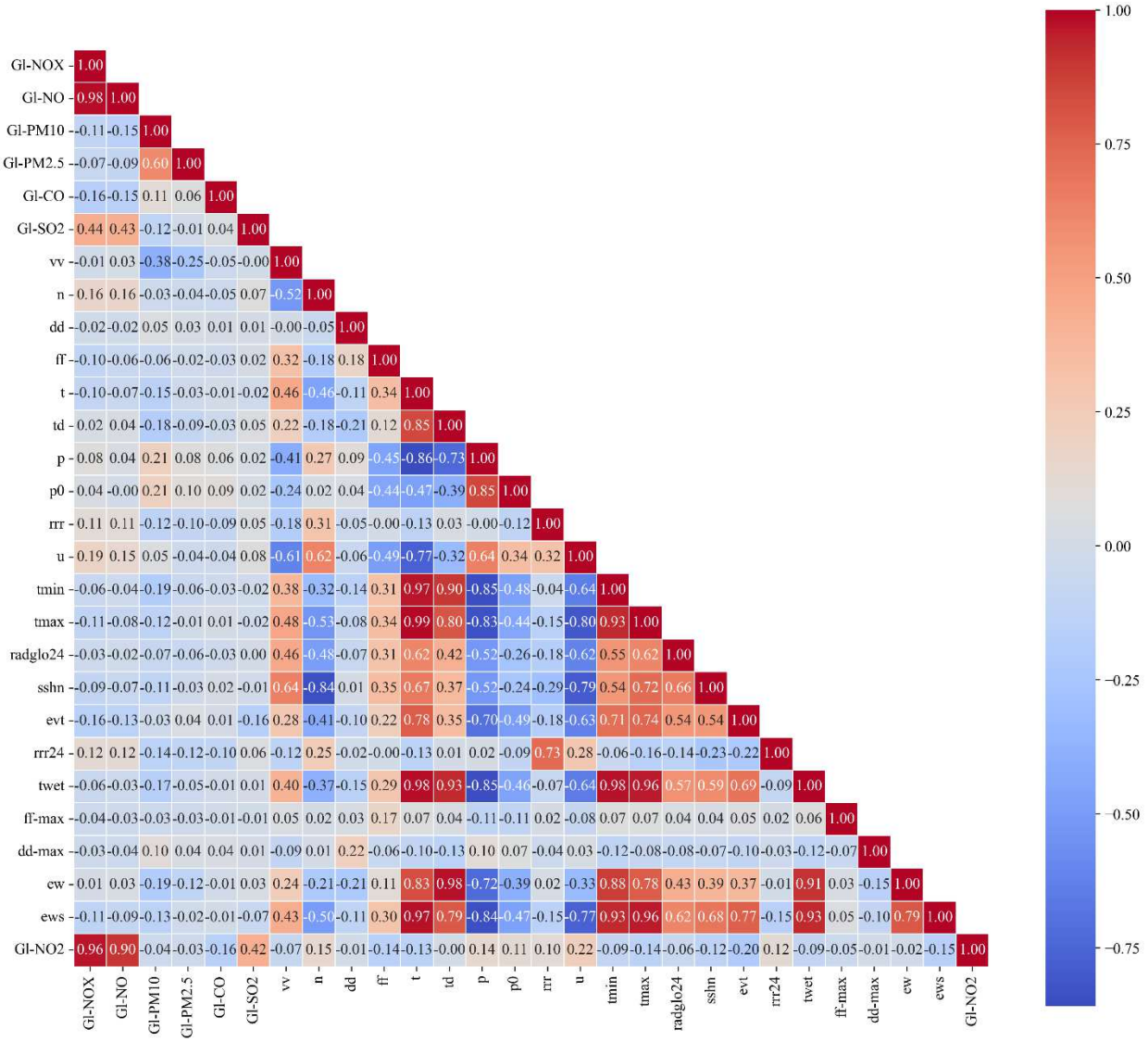


Fig 3- Heatmap of the Correlation Analysis between Ground-Level NO_2 Concentrations and Meteorological Variables and Other Environmental Pollutants

The researchers removed all days with missing data from their study because machine learning models need complete data for accurate results. The study discovered missing information about eight environmental variables which included air temperature and relative humidity and evapotranspiration and wet-bulb temperature and cloud cover and station-level air pressure and ground-based and satellite-derived pollutant concentrations which included NO_2 and NO_X and NO and SO_2.

The researchers processed Sentinel-5P satellite data from Google Earth Engine platform because the system records data only at 13:30 local time by converting all information into daily time intervals which allowed them to synchronize satellite data with ground-based measurements and

meteorological information. The machine learning models operated successfully through multiple dataset usage which created errors from time alignment problems.

The research employed three types of data which included meteorological information and ground-based monitoring records and satellite observations which varied in their numerical measurement ranges. The research team needed to delete incomplete data before they proceeded to transform all remaining data through normalization procedures which maintained identical variable values to stop machine learning models from being affected by different measurement scales. The analysis needed this procedure to establish equal weight for each parameter which resulted in more precise and dependable study outcomes.

The research utilized Sentinel-5P satellite data which included both OFFL (offline) and NRTI (Near-Real-Time Interpolated) products that researchers extracted through Google Earth Engine. The OFFL data requires several days for completion of processing and quality control procedures, while NRTI data enables immediate access for users to perform assessments. The assessment process benefits from using both products because they help to evaluate variations in data quality and accuracy which enables better assessment of ground-level NO_2 concentrations.

The remaining data points for each satellite product were measured after all incomplete records had been eliminated. The final analysis used 667 NRTI data points and 642 OFFL data points from a total of 1860 records which demonstrated a significant level of dependable information.

3.2. Comparative Analysis of Machine Learning Algorithms

The research tested machine learning algorithms which started with basic modeling methods and developed into advanced modeling techniques. The researchers began their analysis with the Decision Tree algorithm to study initial connections between different variables. The researchers chose the Random Forest algorithm because it could decrease overfitting problems and process complicated data sets. The researchers used the Extra Trees algorithm because it produces better model results with increased tree quantity.

The implementation of boosting techniques which include Gradient Boosting and its enhanced version XGBoost resulted in better performance and more precise prediction outcomes. The research utilized CatBoost for categorical data processing and AdaBoost for enhancing weak models to improve forecast accuracy on datasets with high variance.

For model evaluation, the data were split into two subsets: 80% for training and 20% for testing. The coefficient of determination (R^2) root mean square error (RMSE) and mean absolute error (MAE) were used to assess model accuracy and identify the best-performing algorithms. The complete evaluation results are presented in Table 2.

The performance evaluation of various machine learning algorithms for both satellite datasets, NRTI and OFFL, revealed distinct variations in prediction accuracy. The NRTI data showed that the Gradient Boosting algorithm produced the best accuracy results among tested models which

achieved an R^2 value of 0.9961 and a Root Mean Square Error of 0.0578 and a Mean Absolute Error of 0.0212. XGBoost achieved strong results with an R^2 of 0.9946 and RMSE of 0.0680 and MAE of 0.0246. The NRTI data tests showed that AdaBoost produced the lowest accuracy results which reached an R^2 of 0.9880 and RMSE of 0.1021 and MAE of 0.0537. The results demonstrate that boosting algorithms achieve high effectiveness in NO_2 concentration modeling for NRI data through the use of Gradient Boosting and XGBoost methods which maintain accuracy while handling data variability.

Table 2- Comparison of different regression algorithms based on R^2, RMSE, and MAE for NRTI and OFFL datasets.

Machine Learning Algorithms	NRTI			OFFL		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Decision tree	0.9737	0.1517	0.0514	0.9803	0.1300	0.0584
Random Forest	0.9891	0.0973	0.0311	0.9913	0.0923	0.0367
Extra trees	0.9881	0.1168	0.0416	0.9928	0.0849	0.0331
Gradient Boosting	0.9961	0.0578	0.0212	0.9959	0.0634	0.0290
XGBOOST	0.9946	0.0680	0.0246	0.9931	0.0823	0.0370
CATBOOST	0.9900	0.1000	0.0500	0.9900	0.0900	0.0400
AdaBOOST	0.9880	0.1021	0.0537	0.9903	0.0972	0.0545

A similar trend was observed for the OFFL data. The best results were obtained by Gradient Boosting which achieved an R^2 of 0.9959 and an RMSE of 0.0634 and an MAE of 0.0290 while XGBoost followed closely with an R^2 of 0.9931 and an RMSE of 0.0823 and an MAE of 0.0370. The Decision Tree and AdaBoost algorithms produced lower R^2 values together with higher RMSE and MAE results, which demonstrated inferior performance with OFFL data. The results show that boosting algorithms represent the best method to create NO_2 concentration models for OFFL datasets because they can successfully learn the intricate characteristics of processed and quality-controlled data.

The analysis of the two datasets shows that boosting algorithms achieve good performance in both cases yet their accuracy measurements and error rates show small differences when applied to NRTI data which has real-time access and contains increased noise. The OFFL data shows improved accuracy and reduced variability after processing and quality control, which enables Gradient Boosting to better extract data features and make dependable predictions.

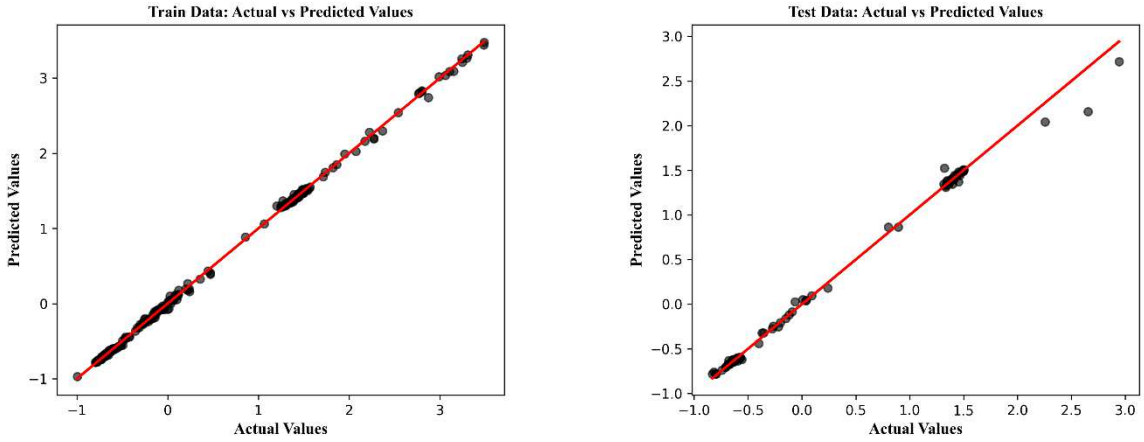
The model accuracy assessment shows that Gradient Boosting performs as the best NRTI and OFFL datasets because it shows lower RMSE values and R^2 values that approach 1. The MAE results support this conclusion because the algorithm delivers predictions that have the smallest absolute error while showing the highest accuracy.

The selected algorithms of this research work together with existing methods to estimate ground-level pollutant concentrations which previous studies have confirmed. The combination of ground-based and satellite data ensures that the models are highly capable of accurately modeling NO_2 concentrations and are suitable for practical applications in air quality monitoring and prediction.

The Gradient Boosting algorithm emerged as the best choice for estimating NO_2 concentrations after performance testing showed that all machine learning models performed different ways. The model evaluation process used three assessment methods which included comparing actual NO_2 measurements with predicted values and analyzing residuals to determine error patterns and creating error distribution histograms. The visualizations demonstrate that the model can precisely replicate the data while validating its capacity to predict NO_2 concentrations accurately in the tested conditions.

The evaluation results of different machine learning algorithms showed that the Gradient Boosting algorithm produced the most precise model for predicting ground-level NO_2 concentrations in this research. The researchers used scatter plots to compare actual NO_2 measurements with model predictions in order to measure the model's performance against real-world data. The researchers performed residual error analysis to study how errors were distributed in the model while they created a histogram to display both prediction error range and error density. The visual analyses show that the model can accurately reproduce real data while it predicts NO_2 concentrations, which functions as a validation method used to assess satellite and ground-based data algorithm performance.

When contextualized against prior literature, these metrics surpass many national and regional-scale studies. For instance, Kim et al. (2021) achieved an R^2 of 0.81 across Europe, Chan et al. (2021) reported an R^2 of 0.79 for Germany, and Long et al. (2022) reported an R^2 of 0.77 for China using similar TROPOMI and meteorological inputs. Even the recent review by Siddique et al. (2024) cites global R^2 values generally ranging between 0.70 and 0.85. The superior performance observed in our study can be attributed to three key methodological distinctions. First, while many large-scale studies rely solely on satellite retrievals and generic meteorological reanalysis (e.g., ERA5), we incorporated co-located, concurrent ground-level concentrations of chemically analogous pollutants—specifically, NOx and NO (which exhibited Pearson correlations of +0.96 and +0.90 with ground-level NO_2, respectively). This provides the algorithm with direct chemical context regarding local combustion sources and atmospheric aging, effectively serving as a strong physical constraint. Second, our study is spatially confined to a single urban point (the monitoring station) rather than a vast heterogeneous grid. This localized modeling reduces spatial noise and allows the algorithm to focus entirely on temporal variability. Third, the rigorous pre-processing—including Z-Score outlier removal, iterative feature selection using absolute Pearson coefficients ($|R| \geq 0.10 |R| \geq 0.10$), and strict temporal matching—ensured that only robust, non-multicollinear signals entered the training phase.



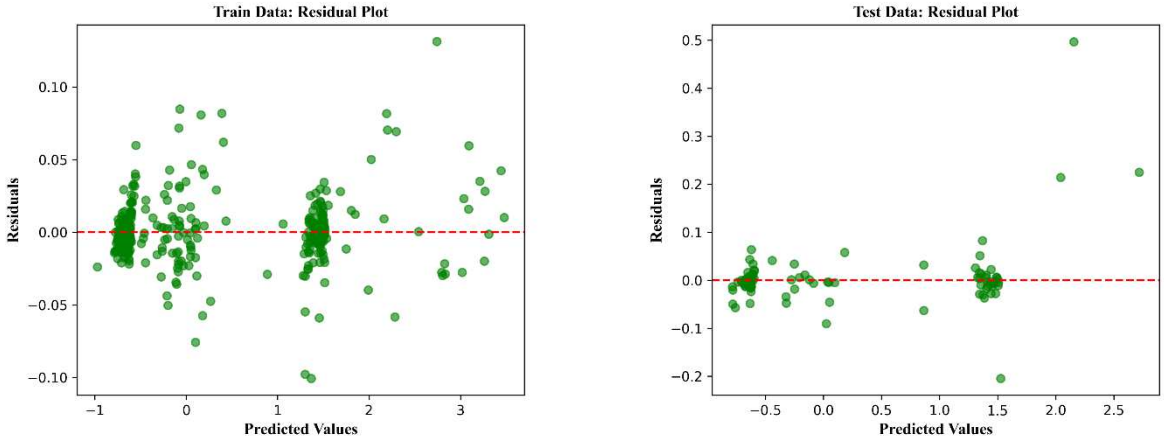
(a) (b)

Fig 4- Comparative scatter plot of observed versus predicted values using the Gradient Boosting algorithm for (a) the training dataset and (b) the test dataset

3.3. Model Validation and Diagnostic Evaluation

The comparative scatter plots of observed versus predicted NO_2 (Figure 4) concentrations using the Gradient Boosting algorithm provide a detailed evaluation of the model's performance for both training and test datasets. In these plots, the x-axis represents the actual measured NO_2 concentrations, while the y-axis corresponds to the concentrations predicted by the model. The training dataset plot shows that the majority of the points are closely aligned along the 45-degree reference line, indicating a high degree of agreement between predicted and actual values. The model demonstrates its capacity to forecast expected results through its precise alignment with the actual data distribution combined with its slight errors and organized pattern of error distribution toward the perfect prediction line.

The scatter plot shows that test data points show people who work for the model which ability to generalize. The data displays greater spread because of higher concentration levels which lead certain data points to move further away from the reference line. The extended distribution of data points shows that the model achieves accurate results through all prediction ranges except when dealing with extreme values. The plots together with residual analysis deliver complete visual evaluation of model accuracy which shows both the capabilities and the limitations of Gradient Boosting to match actual NO_2 levels based on existing data.

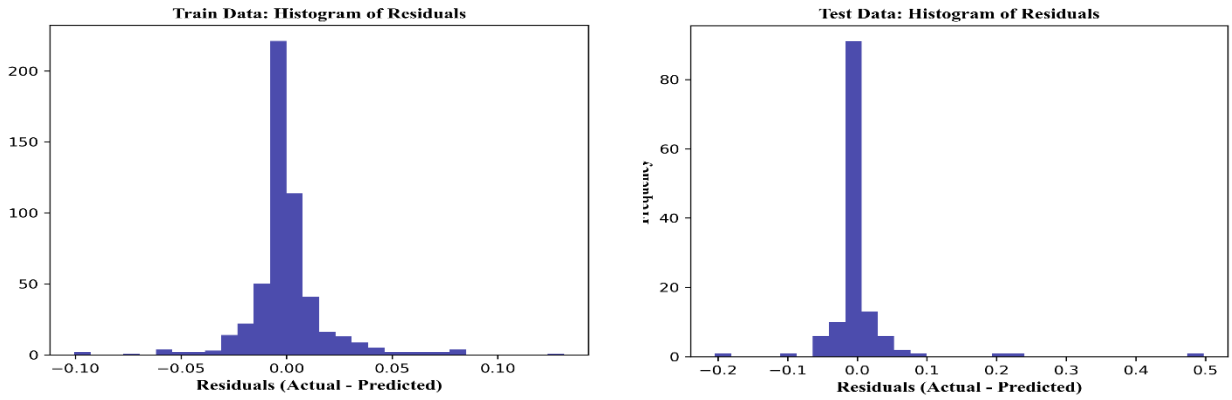


(a) (b)

Fig 5- Residual Analysis Plot for the Gradient Boosting algorithm: (a) Training Dataset and (b) Test Dataset

The analysis of residual plots (Figure 5) uses predicted values to measure their results on the horizontal axis while residuals show the difference between actual results and predicted values on the vertical axis. The Residual Plot shows that most data points cluster around the zero-reference line with symmetrical distribution patterns. The plot shows central clusters which create a point concentration that shows errors maintain stable homoscedasticity across different pollutant concentration levels. The absence of any upward, downward, or U-shaped patterns in the point distribution confirms that the Gradient Boosting model has effectively captured all nonlinear relationships present in the input variables and that the residual errors do not exhibit any systematic bias. The model shows extreme value sensitivity because it only identifies a few outliers who exceed the zero-line threshold while maintaining its overall learning capabilities. The algorithm demonstrates high robustness against uncertainties which affect both ground-level and satellite data because it maintains learning capabilities despite extreme value sensitivity.

The presented figures illustrate the histograms of residual distributions for the Gradient Boosting model across the training (a) and test (b) datasets (Figure 6). The first plot displays residual values which were computed through Actual minus Predicted operations while the second plot shows error frequencies which were measured through vertical axis (Y) of the graph. The training dataset histogram displays a distribution of residuals that shows high concentration at zero which results in an approximately symmetrical distribution pattern. The strong central concentration shows that most training samples produce only minor errors which result in predicted values that match observed measurements. The distribution exhibits a normal distribution pattern because it does not show any significant skewness or high levels of kurtosis which demonstrates that the learning process had no tendency to produce systematic bias and the model appropriately matched the training data.



(a) (b)

Fig 6- Histogram of Gradient Boosting algorithm Error Distribution for (a) Training Dataset and (b) Test Dataset

The test dataset histogram displays its horizontal axis for measuring residual values and its vertical axis for showing their frequency distribution which indicates that residuals extend beyond their training data range. The highest point of error distribution occurs at zero but the model shows greater error dispersion for high residual values which are especially present in the positive section of the distribution. The model shows minor drops in accuracy when assessing extreme cases or situations with high data density. The model demonstrates acceptable statistical stability because it generates error distributions that stay within a small error range without showing multiple peak patterns or strong directional bias and its training and testing performance difference remains consistent. The histograms together with other evaluation metrics show the algorithm achieves correct accuracy without showing overfitting problems while the Gradient Boosting algorithm successfully reconstructs actual pollutant concentrations.

3.4. Environmental Drivers and Regional Implications for Urmia

The modeling framework reveals that Urmia's surface NO_2 concentrations are predominantly governed by local traffic emissions (evidenced by the GL-NOx/NO dominance) and modulated by regional meteorological forcing specific to its semi-arid, high-altitude setting. The proximity to Lake Urmia introduces a unique interplay: as the lake shrinks, exposed saline sediments alter local albedo and heat fluxes, potentially affecting boundary layer stability. Although our model did not explicitly include lake extent, the significant weight assigned to evapotranspiration and relative humidity suggests that moisture recycling from the lake basin plays a subtle yet measurable role in modulating atmospheric mixing.

Operationally, these results provide Urmia's environmental authorities with a scientifically robust, cost-effective framework for continuous monitoring. Given that the Urmia Air Quality Monitoring Station suffers from periodic data blackouts (as seen in 2024), we propose the operationalization of the Gradient Boosting model as an early-warning system. By ingesting freely available Sentinel-5P NRTI data and meteorological forecasts from the Urmia Synoptic Station, authorities can generate daily risk maps. Specifically, during periods of low wind speed, high relative humidity,

and elevated pressure (typical of autumn/winter inversions), the model predicts elevated surface NO_2, which should trigger proactive public health advisories for vulnerable populations (children, elderly, and asthmatics).

3.5. Study Limitations and Future Research Trajectories

Despite the high predictive accuracy, several limitations must be acknowledged to guide future research and ensure transparent application of our results:

The model was trained using a single ground monitoring station (Station No. 3). Consequently, we modeled the temporal dynamics of NO_2 at one specific point in space, not the intra-urban spatial distribution across Urmia (e.g., industrial zones vs. residential suburbs). Future work should deploy mobile monitoring campaigns or low-cost sensor networks to validate satellite-derived columns against multiple surface points, enabling the creation of high-resolution spatial pollution maps using land-use regression (LUR) coupled with this temporal framework.

As noted in Section 3.1, the inclusion of ground-level NOx and NO as predictors creates a dependency on ground infrastructure. While this demonstrates the maximum achievable accuracy, the model's utility in entirely unmonitored areas is limited. Future iterations should explore the use of traffic density proxies (e.g., Google Traffic data or vehicle registration counts) to replace the ground chemical inputs when applying the model to new locations.

The TROPOMI retrieval algorithm relies on air mass factor (AMF) calculations which can be contaminated by mineral dust. Urmia frequently experiences dust events arising from the desiccating Lake Urmia basin. We did not explicitly account for aerosol optical depth (AOD) in our feature set, which may introduce systematic biases during spring dust episodes. Incorporating AOD from the Moderate Resolution Imaging Spectroradiometer (MODIS) could refine future model versions.

The specific meteorological predictors and algorithm hyperparameters identified here are intrinsically tied to Urmia's continental, semi-arid climate and traffic-emission profile. Directly transplanting this model to humid coastal or heavily industrial Iranian cities (e.g., Bandar Abbas or Asaluyeh) without localized retraining is inadvisable. Nevertheless, the methodological pipeline—integrating GEE-extracted TROPOMI data, rigorous correlation-based feature engineering, and boosting algorithms—provides a universal, replicable blueprint for air quality management in data-scarce regions globally.

4. Conclusion

The researchers analyzed ground-based and Sentinel-5P satellite data to measure nitrogen dioxide (NO_2) levels in Urmia city. The ground-based measurements had missing data which decreased the precision of statistical and machine learning models but the combination of satellite data with meteorological information helped to minimize these issues. Pearson correlation analysis showed

that nitrogen oxides (NO_X and NO) most affected ground-level NO_2 while meteorological factors including relative humidity and air temperature and evapotranspiration and cloud cover determined how well ground-based data and satellite-derived data matched.

Modeling with different machine learning algorithms showed that boosting methods, especially Gradient Boosting and XGBoost, produced better NO_2 concentration estimation results. The Gradient Boosting model successfully duplicated the training and testing datasets while maintaining consistent error distribution and demonstrating proper outlier detection capabilities. The model analysis through residual and error histogram tests showed that it functioned without systematic bias and successfully prevented overfitting.

It should be noted that the meteorological parameters and the algorithm yielding the highest accuracy were determined based on the specific geographic, climatic, and data conditions of Urmia. Therefore, the results cannot be directly generalized to other cities or regions. Separate assessments using local data and parameters are necessary to develop accurate and reliable models for each region.

Overall, the study shows that combining satellite data with ground-based data and machine learning algorithms can create accurate models for ground-level NO_2 concentrations which function as dependable tools for monitoring air quality and making environmental decisions in urban areas where ground-based station access is restricted.

Data Availability Statement

The ground-based air quality and meteorological data used in this study are publicly available from the Iran Meteorological Organization and the Urmia Air Quality Monitoring Station (Station No. 3) (supplementary files: gap.xlsx and heatmap.xlsx). Sentinel-5P TROPOMI satellite data were obtained through the Google Earth Engine platform (supplementary files: NTRI.xlsx and OFFL.xlsx). The processed datasets and the Python code used for the machine learning models are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval

Not applicable. This study did not involve human participants, human tissue, or animals.

Consent to Participate

Not applicable. This study did not involve human participants.

Consent to Publish

Not applicable. This study did not involve human participants.

5. References

1- Ahmad, N., Lin, C., Lau, A. K., Kim, J., Zhang, T., Yu, F., ... & Lao, X. Q. (2024). Estimation of ground-level NO_2 and its spatiotemporal variations in China using GEMS measurements and a nested machine learning model. Atmospheric Chemistry and Physics, 24(16), 9645-9665.

2-Ahmady-Birgani, H., Ravan, P., Yao, Z., & Afrasinei, G. M. (2022, May). Newly-Born Sand Dunes of Lake Urmia: Assessing Migration Rate and Morphodynamic Changes Using Remote Sensing Techniques and Field Studies. In EGU General Assembly Conference Abstracts (pp. EGU22-12041). https://doi.org/10.5194/egusphere-egu22-12041

3- Al Yammahi A, Aung Z. Forecasting the concentration of NO_2 using statistical and machine learning methods: A case study in the UAE. Heliyon. 2022 Dec 25;9(2): e12584. doi: 10.1016/j.heliyon. 2022.e12584. PMID: 36793966; PMCID: PMC9922785.

4- Bell, M. L., & Davis, D. L. (2001). Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. Environmental health perspectives, 109(Suppl 3), 389.

5- Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Harmouch, H., & Naumann, F. (2022). The effects of data quality on machine learning performance. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.14529

6- Chan, K. L., Khorsandi, E., Liu, S., Baier, F., & Valks, P. (2021). Estimation of surface NO_2 concentrations over Germany from TROPOMI satellite observations using a machine learning method. Remote Sensing, 13(5), 969.

7- Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., ... & Tao, J. (2022). Machine learning-based estimation of ground-level NO_2 concentrations over China. Science of the Total Environment, 807, 150721.

8- Dou, X., Liao, C., Wang, H., Huang, Y., Tu, Y., Huang, X., ... & Liu, Z. (2020). Estimates of daily ground-level NO_2 concentrations in China based on big data and machine learning approaches. arXiv preprint arXiv:2011.09013.

9- Dubovik, O., Schuster, G. L., Xu, F., Hu, Y., Bösch, H., Landgraf, J., & Li, Z. (2021). Grand challenges in satellite remote sensing. Frontiers in Remote Sensing, 2, 619818.

10- Ghanbarzadeh Lak, Mehdi, rahimzadeh, samin, soleimany, arezoo. “Comparative Analysis of Sulfur Dioxide (SO_2) Concentration with Satellite Data Using Machine Learning Algorithms - A Case Study.” Journal of RS and GIS for Natural Resources, vol. 16, no. 4, 2025, pp. 43-67. https://doi.org/10.82447/girs.2025.1194923. (In Persian).

11- Ghozikali, M. G., Heibati, B., Naddafi, K., Kloog, I., Conti, G. O., Polosa, R., & Ferrante, M. (2016). Evaluation of chronic obstructive pulmonary disease (COPD) attributed to atmospheric O_3, NO_2, and SO_2 using Air Q Model (2011–2012 year). Environmental research, 144, 99-105.

12- Kazemi Garajeh, M., Laneve, G., Rezaei, H., Sadeghnejad, M., Mohamadzadeh, N., & Salmani, B. (2023). Monitoring trends of CO, NO_2, SO_2, and O_3 pollutants using time-series sentinel-5 images based on google earth engine. Pollutants, 3(2), 255-279. https://doi.org/10.3390/pollutants3020019

13- Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., ... & Nieuwenhuijsen, M. (2021). Premature mortality due to air pollution in European cities: a health impact assessment. The Lancet Planetary Health, 5(3), e121-e134.

14- Kim, M., Brunner, D., & Kuhlmann, G. (2021). Importance of satellite observations for high-resolution mapping of near-surface NO_2 by machine learning. Remote Sensing of Environment, 264, 112573.

15- Krotkov, N. A., McLinden, C. A., Li, C., Lamsal, L. N., Celarier, E. A., Marchenko, S. V., ... & Streets, D. G. (2016). Aura OMI observations of regional SO_2 and NO_2 pollution changes from 2005 to 2015. Atmospheric Chemistry and Physics, 16(7), 4605-4629.

16- Long, S., Wei, X., Zhang, F., Zhang, R., Xu, J., Wu, K., ... & Li, W. (2022). Estimating daily ground-level NO_2 concentrations over China based on TROPOMI observations and machine learning approach. Atmospheric Environment, 289, 119310.

17- Mauro, F., Russo, L., Janku, F., Sebastianelli, A., & Ullo, S. L. (2023, October). Estimation of Ground NO_2 Measurements from Sentinel-5P Tropospheric Data through Categorical Boosting. In 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 1116-1121). IEEE.

18- Nouri, A., Ghanbarzadeh Lak, M., & Valizadeh, M. (2021). Prediction of $PM_{2.5}$ concentrations using principal component analysis and artificial neural network techniques: A case study: Urmia, Iran. Environmental Engineering Science, 38(2), 89-98. https://doi.org/10.1089/ees.2020.0089

19- Siddique, Muhammad Adnan & Naseer, Ehtasham & Usama, Muhammad & Basit, Abdul. (2024). Estimation of Surface-Level NO_2 Using Satellite Remote Sensing and Machine Learning: A review. IEEE Geoscience and Remote Sensing Magazine. PP. 2-28. 10.1109/MGRS.2024.3398434.

20- Soleimany, A., Grubliauskas, R., & Šerevičienė, V. (2021). Application of satellite data and GIS services for studying air pollutants in Lithuania (case study: Kaunas city). Air Qual Atmos Health 14: 411–429. https://doi.org/10.1007/s11869-020-00946-z

21- Tavella, R. A., das Neves, D. F., Silveira, G. D. O., Vieira de Azevedo, G. M. G., Brum, R. D. L., Bonifácio, A. D. S., ... & da Silva Júnior, F. M. R. (2025). The Relationship Between Surface Meteorological Variables and Air Pollutants in Simulated Temperature Increase Scenarios in a Medium-Sized Industrial City. Atmosphere, 16(4), 363. https://doi.org/10.3390/atmos16040363

22- Wieczorek, B. (2023). Air pollution patterns mapping of SO_2, NO_2, and CO derived from TROPOMI over Central-East Europe. Remote Sensing, 15(6), 1565.

23- Vineis, P., Demetriou, C. A., & Probst-Hensch, N. (2020). Long-term effects of air pollution: an exposome meet-in-the-middle approach. International journal of public health, 65(2), 125-127.

24- Zhan, Yu & Luo, Yuzhou & Deng, Xunfei & Chen, Huajin & Grieneisen, Michael & Shen, Xueyou & Zhu, Lizhong & Zhang, Minghua. (2017). Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. Atmospheric Environment. 155. 129-139. 10.1016/j.atmosenv.2017.02.023.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The ground-based air quality and meteorological data used in this study are publicly available from the Iran Meteorological Organization (https://data.irimo.ir) and the Urmia Air Quality Monitoring Station (Station No. 3). Sentinel-5P TROPOMI satellite data were obtained through the Google Earth Engine platform (https://earthengine.google.com/). The processed datasets and the Python code used for the machine learning models are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Samin Rahimzadeh: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.

Amir Asadi Vaighan: Conceptualization, Supervision, Project administration, Writing – review & editing, Validation.

Amin Rahimzadeh: Software, Validation, Investigation, Visualization, Data curation.

All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- gap.xlsx
- heatmap.xlsx
- NRTI.xlsx
- OFFL.xlsx