Cross-City Transfer Learning for Sentinel-5P-Driven NO_2 Prediction in Data-Sparse Urban Environments

Fjoralba Janku[1], Francesco Mauro[1], Luigi Russo[2], Babak Memar[3],
Alessandro Sebastianelli[4], Paolo Gamba[2], Silvia Liberata Ullo[1]

[1]Department of Engineering, University of Sannio, Benevento, Italy – f.sota@studenti.unisannio.it;
f.mauro@studenti.unisannio.it; ullo@unisannio.it
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy –
luigi.russo02@universitadipavia.it; paolo.gamba@unipv.it
[3]Sapienza University of Rome, Rome, Italy – babak.memar@uniroma1.it
[4]CMCC Foundation - Euro-Mediterranean Center on Climate Change, Caserta, Italy – asebastianelli@ieee.org

Keywords: NO_2 estimation, Sentinel-5P, transfer learning, remote sensing, CatBoost.

Abstract

Traditional forecasting methods of air pollutants show intrinsic limitations due to the complexity of atmospheric interactions. Recent research has moved toward the employment of artificial intelligence (AI)-based approaches and satellite data processing. The framework proposed in this study is a transfer learning (TL) model to estimate surface-level NO_2 concentrations across multiple locations by using satellite and environmental data. The approach integrates Sentinel-5P TROPOMI-derived tropospheric NO_2 columns, meteorological variables (temperature, precipitation, wind speed and direction), spatial coordinates and temporal features. A CatBoost regression model is implemented, leveraging a Leave-One-City-Out (LOCO) TL framework across five cities (Berlin, London, Madrid, Paris and Toronto) in the world. This enables the model transfer from multiple source domains to a new target city with minimal ground-based data. Experimental results are outperforming city-specific baseline models, by showing a reduced Root Mean Square Error ($RMSE$) by approximately 7% and a Coefficient of Determination (R^2) higher by 2.7%. Toronto, which represents an environment with a low monitoring density, benefits most from TL, with R^2 improving from 0.58 (baseline) to 0.66 (transfer) and RMSE dropping from 6.44 $\mu g/m^3$ to 5.84 $\mu g/m^3$. A detailed Leave-One-Block-Out (LOBO) ablation study shows how each group of features contributes to the performance of the model. Spatial coordinates and meteorological features are the most influential predictors of NO_2 concentration, while the satellite NO_2 data increase model generalization. These results highlight the potential of cross-city TL and remote sensing synergy for scalable urban air pollution monitoring, especially in limited ground-based monitoring scenarios.

1. Introduction

One of the most significant climate-related threats affecting both ecosystem health and human health in recent years is the presence of atmospheric pollutants, including $PM_{2.5}$ and PM_{10} particulate matter, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3). These pollutants are known to contribute to a wide range of adverse effects, such as cardiovascular diseases, premature mortality, and environmental degradation. Among them, nitrogen dioxide (NO_2) stands out as a major pollutant due to its dual impact on health and the environment (World Health Organization, 2022). Furthermore, NO_2 plays a key role in the formation of secondary pollutants, including tropospheric ozone and aerosol nitrates Peng et al. (2022). Despite a recent decline in NO_2 concentrations across Europe, a 2019 report by the European Environment Agency (EEA) indicated that annual NO_2 levels in 22 countries exceeded the European Union (EU) annual limit value of 40 $\mu g/m^3$ (European Environment Agency, 2023). One conventional approach to monitoring air pollutants is through ground-based observation stations. However, these measurements are limited by their point-based nature and heterogeneous spatial distribution Chan et al. (2020). Therefore, it is essential to develop efficient tools for improved monitoring, forecasting, and classification of air quality, thereby enabling more effective and informed decision-making. Traditional forecasting methods rely on physical and chemical models, commonly referred to as Chemical Transport Models (CTMs). For instance, in Menut et al. (2013) , the authors employ such models to simulate pollutant behavior based on meteorological data, emission sources, and chemical reactions. Nevertheless, the application of these models in real-time forecasting remains challenging due to the complexity of atmospheric interactions and the substantial computational resources required. Consequently, recent research has increasingly focused on the use of statistical methods, machine learning, and deep learning techniques for air pollution forecasting, often formulated as a regression task.
Another effective approach to monitoring air pollution is through satellite-based observations, which have significantly advanced in recent years in terms of spatial resolution and data reliability. Earlier missions, such as the Ozone Monitoring Instrument (OMI) Richter and Burrows (2002) and the Global Ozone Monitoring Experiment (GOME) Richter and Burrows (2002), were commonly employed to measure atmospheric NO_2 concentrations. However, the launch of the Sentinel-5P satellite by the European Space Agency (ESA) under the Copernicus Program Veefkind et al. (2012) in October 2017 marked a major breakthrough in air quality monitoring. The satellite's on-board instrument, the TROPOspheric Monitoring Instrument (TROPOMI), provides high-resolution data that enable detailed observation of air pollutants. Several studies have demonstrated the exceptional capability of TROPOMI to capture the spatiotemporal variability of NO_2 concentrations with unprecedented accuracy Shetty et al. (2024), Petetin et al. (2023). Fur-

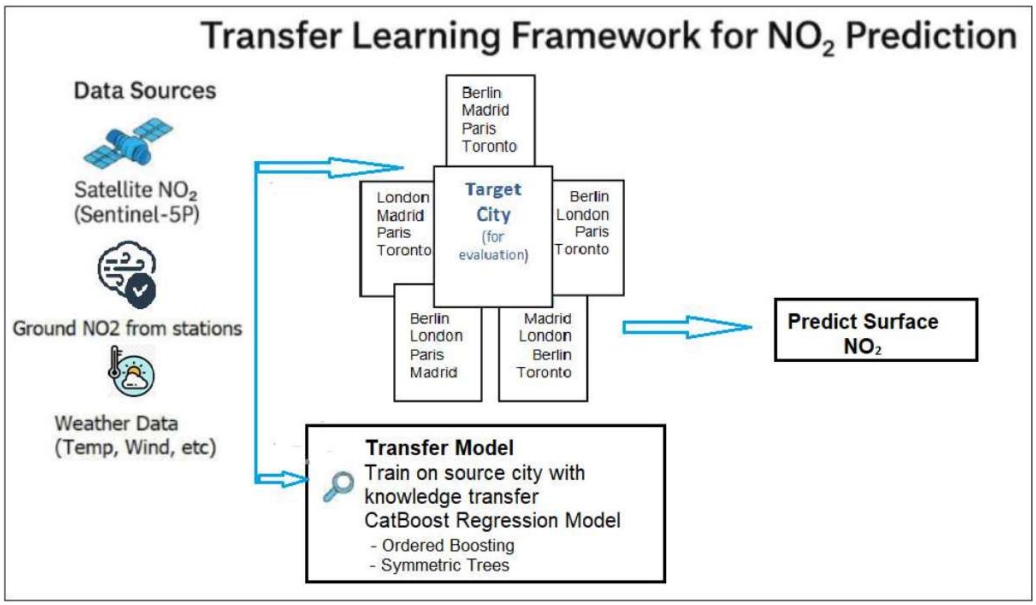

Figure 1. Process of the proposed transfer learning framework in our work.

thermore, the advent of machine learning (ML) and deep learning (DL) techniques over the past decade has significantly improved the accuracy of pollutant estimation compared to traditional statistical approaches. These models are capable of processing large and heterogeneous environmental datasets, demonstrating remarkable performance in air quality prediction tasks. A variety of classical ML models have been employed in this context, including Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost). For instance, in Leong et al. (2020), the authors applied Support Vector Regression (SVR) to forecast the Air Pollution Index. Additionally, studies like Kim et al. (2021) have demonstrated the successful extraction of surface-level NO_2 concentrations from satellite-based datasets using ML frameworks. In a recent study Mauro et al. (2023), the CatBoost was evaluated for its ability to reconstruct the relationship between tropospheric NO_2 concentrations derived from Sentinel-5P data and ground-based measurements across Italy. The objective was to develop a transport model capable of estimating surface emissions when only satellite data were available, thereby enhancing model reliability. Moreover, as highlighted in Jairi et al. (2024), relying on separate models for individual pollutants increases computational costs and prevents the exploitation of shared patterns among different air pollutants. In recent years, transfer learning (TL) has gained considerable attention in the fields of ML and DL for its ability to leverage knowledge from pre-trained models to enhance new tasks. Unlike traditional ML techniques, which require models to be trained from scratch, TL enables the transfer of learned representations from a source model to a target model, effectively reducing computational costs and exploiting similarities between domains Weiss et al. (2016). TL improves model generalization, accuracy, scalability, and interpretability Nowakowski et al. (2025). For instance, Ma et al. (2019) proposed a transferred bi-directional long short-term memory (TL-BLSTM) model for air quality prediction, while Yao et al. (2024) applied a Modified Hybrid Deep Learning model (MHDL) for the estimation of $PM_{2.5}$ in China. More recently, Poelzl et al. (2025) demonstrated that TL techniques improve PM_{10} prediction when transferring a model trained on stations in Graz to Zagreb.

In our study, novel contributions are introduced through the integration of ML techniques with TL and advanced interpretability methods, such as Leave-One-City-Out (LOCO) and Leave-One-Block-Out (LOBO) ablation studies. Namely, we propose a cross-city transfer learning framework for estimating urban NO_2 concentrations, as shown in figure 1, by integrating Sentinel-5P observations, ground-based measurements, and meteorological variables. The approach introduces a LOCO generalization strategy to evaluate cross-regional transferability, a harmonized and multivariate dataset, together with a dedicated feature engineering that enables consistent model training across diverse urban environments, and a LOBO ablation analysis to quantify the contribution of major feature groups. Finally, we demonstrate that the framework is scalable to data-limited cities, requiring only minimal local observations to achieve accurate predictions.
It is worth highlighting that this combined approach, enabling the analysis of the same air pollutant across multiple monitoring stations located in distinct geographic areas, is an aspect that has been only marginally addressed in the existing literature.

The remainder of the paper is organized as follows: Section 2 describes the study areas and data sources. Section 3 describes more in detail the role of TL in air prediction, the TL framework, the proposed methodology, including the preprocessing steps, the CatBoost model setup, and the LOCO validation scheme. Section 4 reports experimental results, comparing baseline and TL models and presenting LOBO analysis for feature importance. Finally, section 5 encapsulates the paper’s findings, providing a concise summary and concluding remarks and prospects.

2. Study areas and Data sources

The study areas have been carefully selected from geographically distant locations, as described in the following section, to demonstrate the proposed method’s scalability and adaptability across diverse urban contexts.

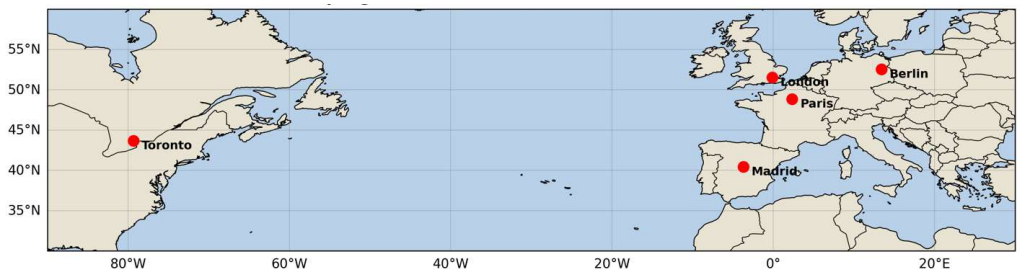

Figure 2. Geographic locations of the five cities considered in this study: Toronto, London, Paris, Berlin and Madrid.

2.1 Study areas

In our study, we selected four major European cities (Berlin, London, Madrid, and Paris) because they represent a wide range of geographic locations, climates, urban structures and sources of pollution. To push the limits of transferability, we further included Toronto in Canada, which differs substantially in geography, climate and emission characteristics. Moreover, the chosen cities offer reliable ground truth (GT) air quality measurements data which are often difficult to obtain consistently, making them ideal testbeds for evaluating the robustness and transferability of predictive models. Figure 2 illustrates the geographic extent of the study area, highlighting the five cities, which are used for model development and evaluation.

2.2 Data sources

A proper harmonized dataset has been created by merging multi-modal raw data from satellite observations, ground monitoring stations, and weather records. This integration enables cross-city learning, enhances data completeness, and supports robust model transferability.

2.2.1 Satellite Observations: Sentinel-5P launched in October 2017 under the European Space Agency (ESA) Copernicus Programme, Sentinel-5P is equipped with the TROPOspheric Monitoring Instrument (TROPOMI), a spectrometer dedicated to global atmospheric composition monitoring. TROPOMI measures multiple trace gases across the ultraviolet (UV), visible (VIS), near-infrared (NIR), and shortwave infrared (SWIR) spectral ranges. Since 2019, it has provided data at a spatial resolution of 3.5×5.5 km. In this study, the offline product is employed, accessed via the Google Earth Engine (GEE) platform during the experimental phase Krotkov et al. (2017). Although Sentinel-5P NO_2 is originally provided as a Level-2 product, the data accessed through GEE are spatially gridded and temporally aggregated, effectively corresponding to a Level-3 representation suitable for machine learning applications.

2.2.2 Ground monitoring stations. Air quality monitoring is performed in the mentioned cities of Berlin, London, Madrid, Paris, and Toronto through their air quality sensor networks. The number of available NO_2 monitoring stations varies across cities, reflecting differences in monitoring infrastructure. These cities are characterized by relatively dense monitoring networks, with several tens of active stations per city, whereas Toronto has a more limited number of stations. This variability makes the selected cities particularly suitable for evaluating transfer learning under heterogeneous data availability. The GT data are collected from publicly available official sources in each city (Greater London Authority, 2024; Umweltbundesamt (UBA), 2024; Ayuntamiento de Madrid, 2024; Airparif, 2024; Ontario Ministry of the Environment, 2024), whose collection is composed of the measurements deriving each from 60, 15, 24, 38 and 4 active monitoring stations, providing hourly ground-level measures of different pollutant concentrations (NO_2, PM_{10}, $PM_{2.5}$, O_3, and SO_2) in $\mu g/m^3$. NO_2 hourly concentrations are downloaded from the 1st of January to the 30th of December 2024, with the related metadata, according to the availability of complete time-series measurements.

2.2.3 Meteorological parameters. Meteorological factors play a crucial role in influencing the concentration and transport of surface-level NO_2. In this study, four key meteorological parameters are considered: temperature (°C), wind direction (°), wind speed (km/h), and precipitation (mm). These data were obtained from the Open-Meteo database (Open-Meteo, 2024), an open-source and freely accessible platform that provides hourly weather information for any global location, with a spatial resolution of approximately 9–10 kilometers.

3. Proposed Methodology

3.1 Transfer Learning for air quality prediction

As previously discussed, TL has emerged as a key approach in modern ML and DL due to its ability to leverage knowledge acquired from previously learned tasks, thereby improving performance and reducing training requirements for related yet data-scarce target tasks. Unlike traditional ML techniques, which require models to be trained from scratch, TL enables the transfer of learned representations from a source model to a target model, effectively reducing computational costs and exploiting similarities between domains Weiss et al. (2016). In this study, novel contributions are introduced through the integration of ML techniques with TL and advanced interpretability methods, such as LOCO and LOBO ablation studies. This combined approach enables the analysis of the same air pollutant across multiple monitoring stations located in distinct geographic areas, an aspect that has been only marginally addressed in the existing literature.

3.2 Model Architecture

The proposed framework employs the CatBoost regression model as the core predictive algorithm due to its superior performance on structured tabular data and its native support for categorical features. CatBoost is based on gradient boosting, building an ensemble of symmetric decision trees, and uses an ordered boosting strategy to reduce overfitting, particularly on small to medium-sized datasets. Its ability to handle categorical variables without the need for manual encoding and its robustness to noisy data make it highly suitable for air quality estimation tasks Hancock and Khoshgoftaar (2020).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-633-2026 | © Author(s) 2026. CC BY 4.0 License.

Two models are compared:

1. **Baseline model:** A CatBoost regression model trained exclusively on data from the target city.

2. **Transfer learning model:** A two-stage model first pre-trained on data from four source cities, followed by fine-tuning using a subset of the target city data. This LOCO TL setup enables knowledge sharing across cities and improves generalization in data-scarce environments.

Evaluation was performed using Root Mean Square Error (RMSE), Normalized RMSE (NRMSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R^2) metrics. Figure 1 sketches the process of the proposed TL approach in our work. To assess model interpretability and quantify the contribution of different data sources, a LOBO ablation study was conducted. This approach systematically removes one group of features at a time, such as satellite-derived variables, meteorological parameters, geographic coordinates, or temporal indicators, to evaluate their individual impact on prediction performance across various urban environments. The resulting performance metrics highlight the relative importance of each feature block, offering valuable insights into the factors that most strongly influence NO_2 prediction within the transfer learning framework.

3.3 Feature engineering

Feature engineering involves extracting and constructing relevant variables from raw data to improve the predictive performance of ML and DL models. In this study, several advanced feature engineering strategies were applied. Beyond the variables obtained from ground measurement stations, additional temporal features were introduced, *month, day_of_week, season, is_weekend*, which influence NO_2 concentrations at the city level Lovrić et al. (2021). Spatial features (*latitude, longitude*) and environmental metrics (*wind_speed, temperature, precipitation*) were also incorporated. For Sentinel-5P data, two input types were generated: one with NO_2 values in the original unit (mol/m^2) and another converted to $\mu g/m^3$ to align with ground-based measurements. The conversion followed the climatological model from Mauro et al. (2023), dividing the tropospheric NO_2 column value by the air column height (h) and multiplying by 1000 and the molar mass of NO_2 (46.055 g/mol). This approximation assumes a homogeneous gas distribution within the column. Finally, both input types were smoothed using the *Savitzky–Golay* filter Schafer (2011), which effectively removes noise while preserving essential signal features.
A detailed description of the dataset columns and corresponding features is provided in Table 1. The training datasets were supplied to the CatBoost algorithm to learn the relationship between the 12 input features listed in Table 1 and the target surface NO_2 measurements obtained from ground stations. As highlighted earlier, one of the new techniques introduced in this study is the LOCO strategy. The CatBoost algorithm was trained using the same hyperparameters as in the previous work, but the dataset was partitioned so that data from the target city were excluded from the training set containing the remaining four cities. The TL model was first pre-trained on all available data from the four source cities and subsequently fine-tuned using a randomly selected 80% subset of the target city's data, with the remaining 20% reserved for testing. The fine-tuning subset was selected via a random stratified split to ensure representative sampling and preservation of the NO_2 concentration distribution.
For both the baseline and TL configurations, the CatBoost regressor was implemented with 10,000 boosting iterations, a learning rate of 0.01, and a maximum tree depth of 12. Early stopping with a patience of 100 iterations was applied to prevent overfitting and optimize computational efficiency. The fine-tuning process allowed the pre-trained model to adapt to the specific characteristics of the target domain, effectively leveraging knowledge transferred from the source domains, in accordance with common TL practices.

3.4 Leave-One-Block-Out (LOBO) Feature Importance Analysis

The LOBO ablation analysis was conducted after the transfer learning process. The model was first pre-trained on the four source cities and subsequently fine-tuned on the target city data following the LOCO strategy. The LOBO analysis was then applied to this trained TL model to assess the contribution of different feature blocks to the final predictive performance. At each step, a specific block of input variables is removed from the training set, and the model is retrained to evaluate the resulting degradation in performance. This method enables us to quantify the importance of high-level feature groups rather than individual variables, offering greater insight into model interpretability. The blocks we consider are: *Remote sensing features*, *meteorological variables*, *spatial location data* and *temporal features*. In our TL setup, we applied LOBO to quantify the importance of a model component on a particular task by measuring the impact of performing ablation on that component. This analysis highlighted which input variables most strongly influenced the transfer-learned model's predictions for each city, providing an interpretable, quantitative summary of feature importance in the CatBoost model. The combination of these two processes has boosted the capability of the entire framework to reach prefixed objectives, with the power of the CatBoost algorithm.
It is worth noting that CatBoost has another significant advantage. It is well-known that the presence of missing values poses a big challenge in ML, as numerous algorithms are unable to accommodate this limitation. On the contrary, CatBoost handles missing values automatically. It learns from the data, and if a feature has missing values, CatBoost does not just throw them away or crash. Instead, it uses a special internal method to treat “missing” as its own value. CatBoost handles categorical variables without needing to manually encode them, so it reduces data preprocessing and prevents overfitting that can result from poor encoding. Moreover, it uses ordered boosting, a permutation-driven approach that avoids target leakage and reduces overfitting, especially on small to medium datasets. In addition, CatBoost is known for being more stable when data are imbalanced or noisy, as in the case of environmental datasets, like for example, when air quality data are employed. In conclusion, CatBoost outperforms several other models, particularly on tabular datasets, because it grows symmetric trees, leading to faster inference and better generalization Hancock and Khoshgoftaar (2020).

4. Results

The two models, the baseline and the TL model, were trained with the same hyperparameters, such as the $number\_of\_iterations = 10000$, $learning\_rate = 0.01$,

Table 1. Description of Dataset Columns and Feature Definitions.

Column Name	Description	Data Type
Latitude	Latitude coordinate of the monitoring station	float64
Longitude	Longitude coordinate of the monitoring station	float64
S5P_smooth	Smoothed Sentinel-5P NO_2 column density	float64
S5P_converted_smooth	S5P converted smooth – Sentinel-5P tropospheric NO_2 column converted from mol/m^2 to µg/m^3 using a climatological air-column approximation and subsequently smoothed using a Savitzky–Golay filter to reduce noise concentration values	float64
month	Month of the observation (1–12)	int64
day_of_week	Day of the week (0 = Monday, ..., 6 = Sunday)	int64
season	Season indicator (1 = Winter, ..., 4 = Autumn)	int64
is_weekend	Weekend indicator (1 = Saturday/Sunday, 0 = Weekday)	int64
temperature_degC	Air temperature in degrees Celsius	float64
precipitation_mm	Precipitation amount in millimeters	float64
wind_speed_kmh	Wind speed in kilometers per hour	float64
wind_direction_deg	Wind direction in degrees (0–360)	int64

$max\_depth = 12$. Height (h) of the gas column contains the tropospheric NO_2 values, by using an intermediate value of 13 km.

4.1 Quantitative analyses

The statistical metrics used in this study to measure the prediction performance of the developed model are the Root Mean Square Error ($RMSE$), the Normalized Root Mean Square Error ($NRMSE$), the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). $RMSE$ is a parameter that measures the difference between the predicted value of the model and the real value. It is very sensitive to extremely large or small errors in a set of data, so it can reflect well the accuracy of the real value. $NRMSE$ scales $RMSE$ to make it dimensionless for easier comparison between datasets or models. Normalizing allows comparison across different scales. Lower $NRMSE$ similarly indicates a better fit (less error variance) relative to the variability of the data. The MAE and (R^2) were also computed to provide complementary insights into model performance. MAE measures the average magnitude of the prediction errors without considering their direction. It's less sensitive to outliers than $RMSE$. R^2 quantifies how much of the variance in the observed data is explained by the model. An R^2 of 1 indicates perfect prediction, while 0 means the model performs no better than the mean. Together, these metrics offer a comprehensive evaluation of predictive accuracy across diverse urban environments.

4.2 Performance Comparison

Two models are compared: *(1) a baseline model* trained only on the target city data and *(2) a TL model* pre-trained on source cities and fine-tuned on the target. Table 2 shows the mean values of *RMSE*, *NRMSE*, *MAE* and R^2 for the five cities taken into consideration. The results demonstrate that the TL approach consistently outperforms the baseline in all evaluation metrics. For example, in Madrid, RMSE decreases from 3.53 to 3.36, MAE from 2.54 to 2.46, and R^2 increases from 0.876 to 0.888. In all cases, TL improves model accuracy, as shown by lower RMSE and MAE values, indicating reduced prediction error; lower NRMSE, showing error reduction relative to data range; higher R^2 scores, suggesting better variance explanation. In particular, Toronto exhibits a notable improvement, with RMSE decreasing from 6.44 to 5.84, NRMSE from 0.1218 to 0.1105, MAE from 4.61 to 4.17, and R^2 increasing from 0.585 to 0.659. This improvement is especially significant given that Toronto has only a few NO_2 monitoring stations, resulting in limited ground-truth data coverage compared to European cities. These consistent gains across multiple evaluation metrics confirm that leveraging knowledge from other cities enhances model performance, especially when the target city has limited or noisier data.

4.3 Prediction Accuracy Visualization

Figure 3 presents a comparative visualization of predicted versus observed ground-level NO_2 concentrations for five major cities. Each subplot consists of two panels: the left panel displays the sequence from 1 January to 30 December 2024 of NO_2 predictions, while the right panel zooms in on a representative block of 25 contiguous days to highlight finer temporal dynamics. Across all cities, the TL model (orange dashed line) consistently aligns more closely with the true NO_2 values (black line) than the baseline model (blue dashed line). This visual trend confirms the quantitative improvements observed in RMSE and NRMSE metrics. In particular, in cities like Madrid and Paris, the TL predictions exhibit better tracking of the peak values and overall variability in pollution levels, indicating that knowledge transferred from other cities enhances the model's generalization capabilities. This enhanced tracking is visually evident in the zoomed panels, where the TL model's trajectory nearly overlays the ground truth.

Figure 4 presents scatter plots comparing predicted versus true NO_2 concentrations across five cities using both the baseline and TL models. In each plot, predictions from the baseline model are shown in green, while those from the TL model appear in red. The diagonal line ($y = x$) indicates perfect prediction.

Across all cities, the TL model yields predictions more closely aligned with the true NO_2 values, as evidenced by the red points clustering more tightly around the identity line. This improved alignment is supported by consistently higher R^2 values for the TL model, with notable gains in cities like Toronto ($R^2$from 0.58 to 0.66) and London ($R^2$from 0.86 to 0.88). These results confirm that incorporating knowledge from other cities enhances predictive accuracy, particularly in data-sparse environments.

4.4 LOBO Analysis for NO_2 Prediction Across Cities

To interpret the contribution of different input feature groups to model performance, we apply a LOBO ablation analysis across five target cities. LOBO isolates feature group impact without retraining from scratch.

For each target city, models were trained excluding one feature block at a time: satellite data, meteorological variables, spatial coordinates, or temporal features, while evaluating performance using $RMSE$, $NRMSE$, MAE, and R^2. Every other training detail is held constant while ablating a block. The results

Table 2. Performance comparison between baseline and TL models across different target cities.

Target City	RMSE (Base)	NRMSE (Base)	MAE (Base)	R^2 (Base)	RMSE (Transfer)	NRMSE (Transfer)	MAE (Transfer)	R^2 (Transfer)
Berlin	8.13	0.0856	6.18	0.798	7.96	0.0837	5.98	0.807
London	4.68	0.0505	3.12	0.865	4.41	0.0476	2.99	0.880
Madrid	3.53	0.0560	2.54	0.876	3.36	0.0533	2.46	0.888
Paris	4.12	0.0598	2.86	0.886	3.83	0.0556	2.71	0.902
Toronto	6.44	0.1218	4.61	0.585	5.84	0.1105	4.17	0.659

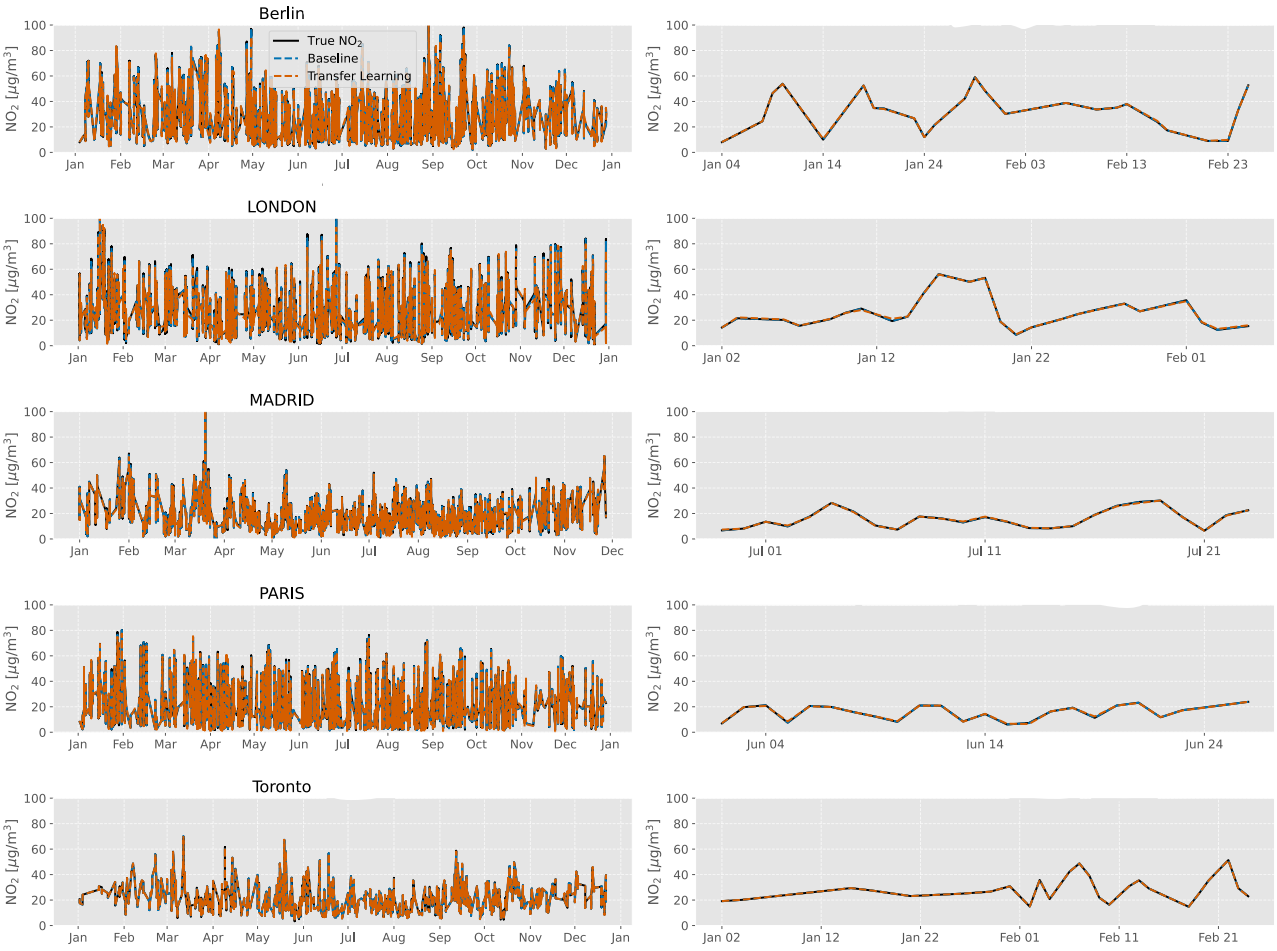

Figure 3. Predicted and observed ground-level NO_2 concentrations for five cities, comparing baseline and TL models. The left panels show the sequence from 1 January to 30 December 2024; the right panels present a zoomed view of the first 25 contiguous daily samples for each city.

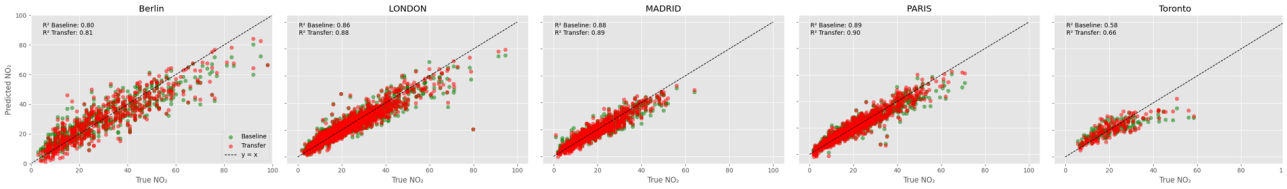

Figure 4. Comparison of NO_2 Predictions: Baseline vs Transfer Learning.

in Table 3 demonstrate that removing spatial coordinates consistently led to the most significant performance degradation across all cities. For example, in Berlin, excluding latitude and longitude increases RMSE to 15.23 and reduces R^2 to 0.29, compared to substantially better performance when other feature blocks are removed. Similar behavior is observed in London, where R^2 drops sharply to 0.32, and in Paris, where it decreases to 0.14. These results highlight the critical role of spatial coordinates in capturing persistent intra-urban emission patterns, such as traffic corridors, industrial zones, and land-use characteristics. Meteorological variables emerge as highly influential feature group. Their removal consistently leads to large increases in error and marked reductions in R^2 across all cities. For instance, in Madrid, removing meteorological information increases RMSE to 5.58 and reduces R^2 to 0.69, while in Toronto R^2 drops to 0.42. This behavior reflects the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-633-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 3. LOBO Ablation Study: Performance Impact of Feature Removal on TL for NO_2 Prediction Across Cities

Target City	Ablation Set	RMSE	NRMSE	MAE	R^2
Berlin	No_Satellite	7.9962	0.0802	5.6225	0.8228
	No_Meteo	12.4054	0.1306	9.0456	0.5300
	No_Coordinates	15.2324	0.1603	12.1234	0.2914
	No_CoreTime	8.2292	0.0866	6.2266	0.7932
London	No_Satellite	4.4800	0.0462	2.8706	0.8868
	No_Meteo	7.5437	0.0814	5.3772	0.6482
	No_Coordinates	10.5139	0.1134	7.5315	0.3167
	No_CoreTime	4.9091	0.0530	3.4033	0.8510
Madrid	No_Satellite	3.4789	0.0520	2.3949	0.8929
	No_Meteo	5.5812	0.0886	4.1131	0.6896
	No_Coordinates	6.6248	0.1052	4.9671	0.5626
	No_CoreTime	4.0137	0.0637	2.9704	0.8395
Paris	No_Satellite	3.9868	0.0534	2.5851	0.9090
	No_Meteo	5.9462	0.0862	4.3520	0.7634
	No_Coordinates	11.3547	0.1646	8.2311	0.1373
	No_CoreTime	4.4787	0.0649	3.1811	0.8658
Toronto	No_Satellite	5.8024	0.1103	4.1943	0.6517
	No_Meteo	7.6357	0.1444	5.5175	0.4167
	No_Coordinates	9.0649	0.1715	6.6770	0.1779
	No_CoreTime	6.7677	0.1280	4.7295	0.5418

strong dependence of NO_2 concentrations on atmospheric processes governing dispersion, accumulation, and transport, including wind speed, temperature, and precipitation. Temporal features contribute more moderately but consistently across all cities. Excluding temporal information leads to small yet systematic increases in error and reductions in R^2, indicating that seasonal and weekly patterns help stabilize predictions but are secondary to spatial and meteorological drivers. For example, in London, removing temporal features reduces R^2 from values close to 0.88 to 0.85, while in Paris it decreases to 0.87. Satellite-derived NO_2 features exhibit a relatively smaller marginal impact in the LOBO analysis. In several well-monitored cities, such as Berlin, Madrid, and Paris, removing satellite features results in only minor changes in RMSE and R^2. However, in Toronto, characterized by a sparser monitoring network—the removal of satellite features leads to a noticeable degradation in performance, underscoring their importance in data-scarce environments. In some cases, the removal of a feature block results in a increase in R^2. This behavior is attribute of feature redundancy and noise effects among correlated predictors. When a block contains partially redundant or noisy information (e.g., satellite observations in well-monitored cities), its removal may reduce overfitting and slightly improve variance explanation. This does not indicate that the removed features are unimportant, but rather reflects the regularization behavior of CatBoost. Importantly, spatial and meteorological features consistently cause the largest performance degradation when removed, confirming their dominant predictive role. At first glance, it may seem surprising that spatial coordinates emerge as the most influential predictors, while satellite-derived NO_2 appears to contribute less in the LOBO analysis. This behavior reflects the strong spatial structure of urban NO_2 pollution. Latitude and longitude implicitly capture a wide range of persistent urban characteristics, such as traffic intensity, emission hotspots, land-use patterns, and proximity to city centers. As a result, spatial coordinates are particularly effective at representing stable, fine-scale gradients that are difficult to fully resolve using satellite observations alone. Satellite-derived NO_2 from Sentinel-5P provides valuable information on broader spatial patterns and background pollution levels, but its coarser spatial resolution means that part of its signal overlaps with information already captured by spatial and meteorological variables. Within the LOBO framework, which evaluates the additional contribution of each feature group in the presence of all others, this overlap can lead to a smaller apparent impact of satellite features, especially in cities with dense ground monitoring networks. This outcome does not imply that satellite data are of limited value. On the contrary, satellite observations are a key component of the proposed transfer learning framework, as they provide consistent, spatially continuous information across cities. Their importance becomes particularly evident in data-scarce environments, where they help bridge gaps in ground-based measurements, as demonstrated by the substantial performance improvements observed for Toronto. Overall, satellite features primarily support model generalization and transferability across heterogeneous urban contexts, rather than dominating local predictions when rich ground-level information is available.

5. Conclusions

This study presents a transfer learning (TL) framework enhanced with LOBO ablation-based explainability for predicting urban NO_2 concentrations in data-scarce environments. A harmonized multi-city dataset was constructed for London, Madrid, Paris, Berlin, and Toronto by integrating Sentinel-5P satellite observations, ground-based measurements, meteorological variables, and spatiotemporal features. Using a Leave-One-City-Out (LOCO) strategy, the proposed TL model—pre-trained on multiple source cities and fine-tuned using limited target-city data—consistently outperformed baseline models trained solely on local observations. Performance improvements were observed across all cities, demonstrating that knowledge learned from data-rich urban environments can effectively support air quality prediction where monitoring data are limited. The results obtained for Toronto, where R^2 improved from 0.585 to 0.659, particularly highlight the robustness of the approach under sparse monitoring conditions and emphasize the importance of satellite observations in compensating for missing ground measurements. The LOBO ablation analysis further provided interpretability by quantifying the contribution of feature groups. Spatial coordinates emerged as the most influential predictors, followed by meteorological variables, while temporal features contributed to prediction stability. Although satellite-derived NO_2 showed a smaller marginal contribution in well-monitored cities, it remains essential for improving model transferability and scalability across heterogeneous urban environments. Together, the LOCO and LOBO analyses demonstrate both the generalization capability of the TL framework and the relative importance of multi-source predictors. Despite promising results, several limitations remain. The current framework does not explicitly incorporate land-use or population-related indicators, and evaluation was limited to a single machine learning architecture. Future work should explore additional ML and deep learning approaches, integrate higher-resolution satellite data, and extend the framework toward real-time urban air quality applications. Model performance during high-pollution episodes remains challenging due to the limited representation of extreme events in training data. Nevertheless, such episodes were retained during preprocessing and the model successfully reproduced their temporal evolution. Future research should explicitly address extreme pollution conditions through dedicated modeling strategies and expand the approach to additional pollutants and urban regions.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-633-2026 | © Author(s) 2026. CC BY 4.0 License.

References

Airparif, 2024. Paris air quality monitoring data (open data). https://data-airparif-asso.opendata.arcgis.com/. Accessed: 2024-10-20.

Ayuntamiento de Madrid, 2024. Calidad del aire - madrid air quality stations. https://www.madrid.es/portal/site/munimadrid. Accessed: 2024-10-20.

Chan, K. L., Wiegner, M., van Geffen, J., De Smedt, I., Alberti, C., Cheng, Z., Ye, S., Wenig, M., 2020. MAX-DOAS measurements of tropospheric NO 2 and HCHO in Munich and the comparison to OMI and TROPOMI satellite observations. *Atmospheric Measurement Techniques*, 13(8), 4499–4520.

European Environment Agency, 2023. Sources and emissions. https://www.eea.europa.eu/en/analysis/publications. Accessed October 21, 2025.

Greater London Authority, 2024. Air quality monitoring sites dataset. https://data.london.gov.uk/dataset/air_quality_monitoring_sites. Accessed: 2024-10-20.

Hancock, J. T., Khoshgoftaar, T. M., 2020. CatBoost for big data: an interdisciplinary review. *Journal of big data*, 7(1), 94.

Jairi, I., Ben-Othman, S., Canivet, L., Zgaya-Biau, H., 2024. Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants. *Environmental Technology & Innovation*, 36, 103793.

Kim, M., Brunner, D., Kuhlmann, G., 2021. Importance of satellite observations for high-resolution mapping of near-surface NO2 by machine learning. *Remote Sensing of Environment*, 264, 112573.

Krotkov, N. A., Lamsal, L. N., Celarier, E. A., Swartz, W. H., Marchenko, S. V., Bucsela, E. J., Chan, K. L., Wenig, M., Zara, M., 2017. The version 3 OMI NO_2 standard product. *Atmospheric Measurement Techniques*, 10(9), 3133–3149. https://amt.copernicus.org/articles/10/3133/2017/.

Leong, W., Kelani, R., Ahmad, Z., 2020. Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, 8(3), 103208.

Lovrić, M., Pavlović, K., Vuković, M., Grange, S. K., Haberl, M., Kern, R., 2021. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. *Environmental pollution*, 274, 115900.

Ma, J., Cheng, J. C., Lin, C., Tan, Y., Zhang, J., 2019. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment*, 214, 116885.

Mauro, F., Russo, L., Authors withheld, Sebastianelli, A., Ullo, S. L., 2023. Estimation of ground no2 measurements ...

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., Coll, I., Curci, G., Foret, G., Hodzic, A. et al., 2013. CHIMERE 2013: a model for regional atmospheric composition modelling. *Geoscientific model development*, 6(4), 981–1028.

Nowakowski, A., Rosso, M. P. D., Zachar, P., Spiller, D., Gabara, G., Barretta, D., Kalinowska, K. B., Choromański, K., Wilkowski, A., Sebastianelli, A., Kupidura, P., Osińska-Skotak, K., Ullo, S. L., 2025. Transfer Learning in Earth Observation Data Analysis: A review. *IEEE Geoscience and Remote Sensing Magazine*, 13(1), 121-152.

Ontario Ministry of the Environment, 2024. Air quality ontario - monitoring network. https://www.airqualityontario.com/. Accessed: 2024-10-20.

Open-Meteo, 2024. Historical weather api documentation. https://open-meteo.com/en/docs/historical-weather-api. Accessed: 2024-10-21.

Peng, S., Lin, X., Thompson, R. L., Xi, Y., Liu, G., Hauglustaine, D., Lan, X., Poulter, B., Ramonet, M., Saunois, M. et al., 2022. Wetland emission and atmospheric sink changes explain methane growth in 2020. *Nature*, 612(7940), 477–482.

Petetin, H., Guevara, M., Compernolle, S., Bowdalo, D., Bretonnière, P.-A., Enciso, S., Jorba, O., Lopez, F., Soret, A., Pérez García-Pando, C., 2023. Potential of TROPOMI for understanding spatio-temporal variations in surface NO 2 and their dependencies upon land use over the Iberian Peninsula. *Atmospheric Chemistry and Physics*, 23(7), 3905–3935.

Poelzl, M., Kern, R., Kecorius, S., Lovrić, M., 2025. Exploration of transfer learning techniques for the prediction of PM10. *Scientific Reports*, 15(1), 2919.

Richter, A., Burrows, J., 2002. Tropospheric NO2 from GOME measurements. *Advances in Space Research*, 29(11), 1673-1683. https://www.sciencedirect.com/science/article/pii/S027311770200100X.

Schafer, R. W., 2011. What is a Savitzky-Golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4), 111–117.

Shetty, S., Schneider, P., Stebel, K., Hamer, P. D., Kylling, A., Berntsen, T. K., 2024. Estimating surface NO2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning. *Remote Sensing of Environment*, 312, 114321.

Umweltbundesamt (UBA), 2024. Luftdaten stationen – air quality data for berlin. https://www.umweltbundesamt.de/daten/luft/luftdaten/stationen. Accessed: 2024-10-20.

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., De Vries, J., Otter, G., Claas, J., Eskes, H., De Haan, J., Kleipool, Q. 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 sensing of environment*, 120, 70–83.

Weiss, K., Khoshgoftaar, T. M., Wang, D., 2016. A survey of transfer learning. *Journal of Big data*, 3, 1–40.

World Health Organization, 2022. Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1. Accessed October 21, 2025.

Yao, B., Ling, G., Liu, F., Ge, M.-F., 2024. Multi-source variational mode transfer learning for enhanced PM2. 5 concentration forecasting at data-limited monitoring stations. *Expert Systems with Applications*, 238, 121714.