Environmental Pollution 383 (2025) 126691

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Using open data to derive parsimonious data-driven models for uncovering the influence of local traffic and meteorology on air quality: The case of Madrid

Koorosh Kazemi *, Anton Vernet, Alexandre Fabregat

Department of Mechanical Engineering, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007, Tarragona, Spain

ARTICLE INFO

Dataset link: www.datos.madrid.es

Keywords:
Air pollution
Machine learning
Road traffic emissions
Urban air quality

ABSTRACT

Air pollution remains a critical public health and environmental challenge, particularly in urban areas where traffic emissions and meteorological conditions strongly influence air quality. While Machine Learning (ML) techniques have been increasingly used to model pollutant concentrations, many existing studies rely on complex architectures that often integrate multiple heterogeneous data sources. In contrast, this study presents a parsimonious, data-driven ML model that predicts local hourly concentrations of key pollutants—NO_2, O_3, $PM_{2.5}$, and PM_{10}—in Madrid using only open data sources. A key factor of our approach is the incorporation of hourly road traffic data collected in the immediate vicinity of each pollutant monitoring station as a predictor. This localized traffic information, combined with local meteorological data, allows our model to outperform other existing solutions that often depend on historical and/or proprietary data. Our results clearly demonstrate that better data might surpass the benefits of more complex ML architectures. The model achieves strong predictive accuracy, with test R^2 scores ranging from 0.77 to 0.86 for NO_2, 0.8 to 0.85 for O_3, 0.63 to 0.82 for $PM_{2.5}$, and 0.68 to 0.95 for PM_{10}. This remarkable performance underscores the utility of dense networks of vehicle count sensors providing high-resolution spatiotemporal traffic data as a critical input for accurate urban air quality modeling. Additionally, we conducted a sensitivity analysis to assess the impact of reducing vehicle emissions on local NO_2 levels, offering actionable insights for policymakers. The findings highlight the potential of open-data-driven models in urban air quality management, providing a scalable, cost-effective, and interpretable tool to support evidence-based decision-making and environmental policy design.

1. Introduction

Air pollution presents a pressing global challenge, threatening human health, ecosystems, and environmental sustainability at an unprecedented scale (Manisalidis et al., 2020). This issue has been exacerbated by rapid industrialization and urbanization, with pollutants emanating from diverse sources such as industrial processes (García-Pérez et al., 2007), vehicular emissions (Han and Naeher, 2006), and natural phenomena (Gonçalves et al., 2009). The resulting air pollutants—including particulate matter (PM), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3)—are linked to severe health conditions such as respiratory diseases, cardiovascular problems, and premature mortality (Kim et al., 2018; Lee et al., 2014; Lelieveld et al., 2015). These contaminants also contribute to ecosystem degradation, climate change, and a decline in air quality standards (Karnosky et al., 2003; Ramanathan and Feng, 2009; Akimoto, 2003). Addressing this multifaceted challenge is especially critical in urban areas, where the interplay of population density, traffic, and industrial activity create complex environmental dynamics (Krzyzanowski et al., 2014).

Madrid, Spain's capital and one of Europe's major urban hubs, faces significant air quality challenges. The city's dense road network, heavy traffic and distinct climatic conditions exacerbate pollution levels, affecting the health of its residents and the resilience of its ecosystems (Laña et al., 2016; Izquierdo et al., 2020). In recent years, episodes of elevated air pollution have heightened public awareness and underscored the need for effective mitigation strategies (Silva et al., 2017; Wang et al., 2022). Predicting air quality in Madrid has become a crucial component of urban planning, allowing for informed decision-making and the implementation of proactive measures to safeguard public health and the environment (Carmichael et al., 2008).

Numerous predictive models have been developed to estimate air pollutant concentrations, employing a wide range of methods, from traditional statistical techniques

* Corresponding author.
E-mail addresses: koorosh.kazemi@urv.cat (K. Kazemi), anton.vernet@urv.cat (A. Vernet), alexandre.fabregat@urv.cat (A. Fabregat).

https://doi.org/10.1016/j.envpol.2025.126691
Received 26 February 2025; Received in revised form 6 June 2025; Accepted 16 June 2025
Available online 5 July 2025

(Chelani et al., 2001; Zolghadri and Henry, 2004; Hassanzadeh et al., 2009; Kumar and Jain, 2010; Gocheva-Ilieva et al., 2014; Jaiswal et al., 2018) to advanced machine learning and deep learning algorithms (Ayturan et al., 2018; Lin et al., 2025; Ansari and Quaff, 2025; Li et al., 2025). While state-of-the-art models such as neural networks and ensemble methods offer high accuracy, they often come with significant limitations. On the one hand, they rely on complex architectures, require large amounts of preprocessed data, and frequently incorporate proprietary datasets. In addition, many architectures use as predictors previous pollutant measurements that often emerge as the most important feature in predicting the current value of concentration (Liang et al., 2020; Wang et al., 2020). All in all, these approaches not only require important computational resources but also introduce challenges in interpretability, making it difficult to disentangle the relative contributions of key factors such as meteorology and road traffic to pollution levels.

Finally, these methodologies are often constrained by their dependence on large volumes of heterogeneous third-party data and by substantial computational demands.

This study adopts a parsimonious approach to air quality modeling, relying exclusively on publicly available open data and established methodologies. Unlike many predictive models, ours does not depend on past pollutant concentration data, enabling a direct assessment of how road traffic and meteorological conditions influence air pollution levels. By eliminating the need for complex preprocessing pipelines, this approach enhances interpretability while maintaining strong predictive performance, as evidenced by the high R^2 values achieved.

A key strength of this method lies in its use of highly localized data on road traffic intensity and meteorology to predict pollutant concentrations at the precise locations where they are measured. Although the model is inherently limited to predicting air quality at existing monitoring stations, the dense spatial distribution of sensors across Madrid, coupled with the model's strong predictive performance, supports its use as a practical tool for spatial air quality analysis and planning, offering a simpler yet effective alternative to more complex modeling approaches. By offering a scalable, transparent, and accessible solution, this approach facilitates broader adoption and provides valuable insights into the key drivers of air pollution in Madrid.

The feature importance analysis highlights the crucial role of meteorological factors – including temperature, wind speed, and humidity – in the dispersion and transformation of pollutants. At the same time, road traffic remains the primary driver of localized pollution hotspots. Understanding this interplay is essential for tackling air quality challenges effectively. Unlike many previous studies that incorporate temporal pollutant trends as predictive variables (Zhu et al., 2018; Castelli et al., 2020; de Medrano et al., 2021), our model isolates these effects by relying exclusively on real-time meteorological and traffic data.

2. Methodology and data

2.1. Problem definition

This study presents a Machine Learning based model designed to predict hourly air pollutant concentrations while maintaining a parsimonious approach regarding predictor variables. Predicting hourly concentrations refers to estimating pollutant levels for a given hour (t) using predictor variables such as meteorological conditions, traffic intensity, and power consumption measured at the same hour (t). The model does not incorporate time-lagged inputs or historical pollutant data, and is therefore not intended for traditional time series forecasting. Instead, it provides hour-specific concentration estimates based on contemporaneous conditions, supporting real-time or scenario-based applications. The model exclusively relies on open data, ensuring both accessibility and reproducibility. Specifically, it integrates very local meteorological data and road traffic intensity as key inputs, both critical factors influencing air quality. By relying on a small set of predictors and accessible public datasets, the model provides an efficient yet powerful solution for air quality prediction while simultaneously assessing the importance of each variable. This approach improves transparency while also lowering computational demands, allowing it to be more easily applied in diverse areas such as environmental monitoring and public health.

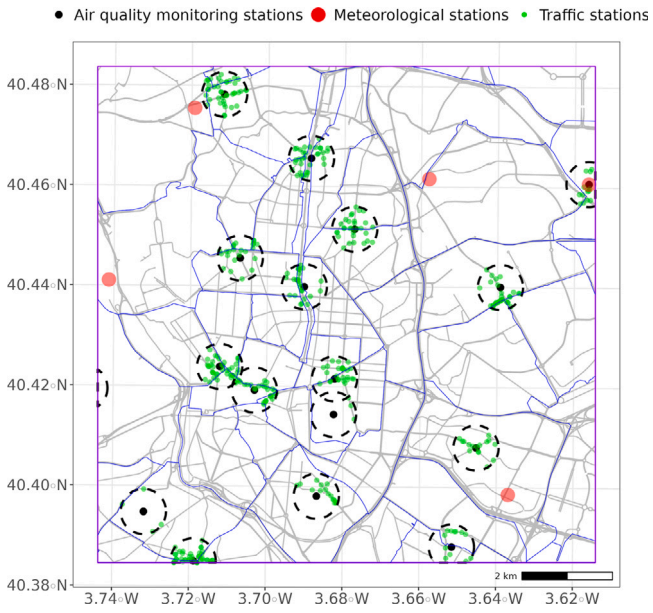

Fig. 1. Map of central Madrid showing the locations of air pollutant monitoring stations (black markers), their surrounding vicinity areas (black dashed circles), and the road traffic sensors located within these areas (green markers). Meteorological stations are indicated in red, and zip code boundaries are shown in blue. Major roads are depicted as thin black lines.

2.2. Data overview

All data for this research was sourced from the official data portal of Madrid (www.datos.madrid.es). The dataset comprises data collected from January 2022 through December 2023 (24 months).

2.2.1. Pollutant concentration

This dataset includes hourly measurements of pollutant concentrations recorded at the 22 measuring stations across Madrid. Along with the major street network in the central area of the city, Fig. 1 shows the location of 16 out of 22 pollutant measuring stations as black markers. The full list of stations, including their ID, name, location (latitude and longitude), and address, is shown in Table A.1 in Appendix A (see also Fig. A.1 for a complete map).

These 22 monitoring stations record concentrations of various pollutants, including carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NOx), ozone (O_3), particulate matter with a diameter of 10 micrometers or less (PM_{10}), particulate matter with a diameter of 2.5 micrometers or less ($PM_{2.5}$), and sulfur dioxide (SO_2). In this study, we focus on the pollutants most commonly available across all stations, namely NO_2, O_3, $PM_{2.5}$, and PM_{10}. These pollutants are of particular concern due to their impact on human health. Thus, according to the World Health Organization (WHO), nitrogen dioxide (NO_2) is a key urban air quality concern, primarily due to its prevalence and contribution to respiratory issues. Particulate matter, especially $PM_{2.5}$, is considered the most hazardous to human health because it can penetrate deep into the lungs and even enter the bloodstream (Fazakas et al., 2024).

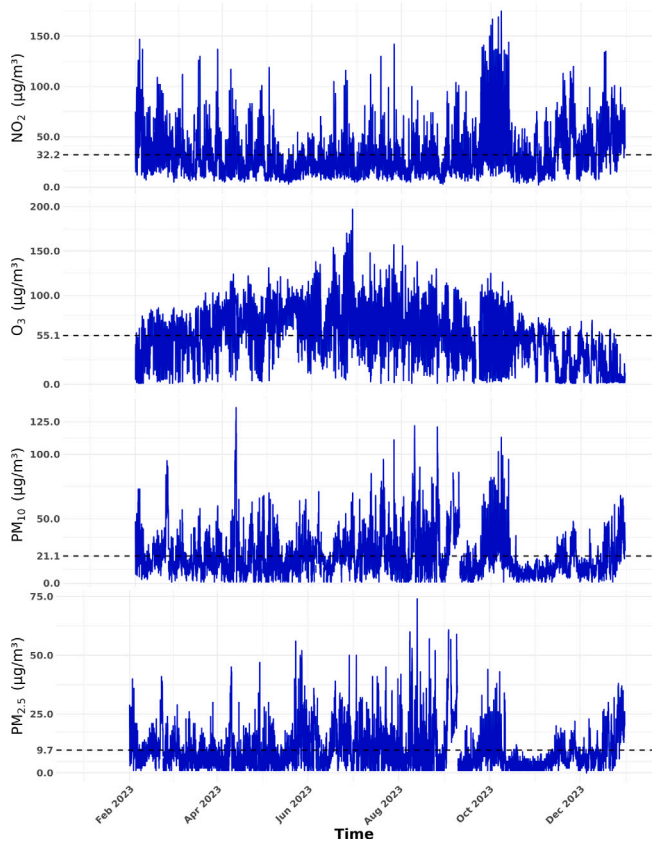

Fig. 2. Hourly time series of different pollutants at location = 8 for 2023; the black dashed line shows the average value of each pollutant.

For illustration, Fig. 2 presents the time series of pollutant concentrations (in $\mu g\,m^{-3}$) at an arbitrary location (measuring station 8). The four panels, from top to bottom, correspond to NO_2, O_3, PM_{10}, and $PM_{2.5}$. The dashed black line in each panel represents the annual mean concentration, with values of 32.2, 55.1, 21.1, and 9.7 $\mu g\,m^{-3}$ for NO_2, O_3, PM_{10}, and $PM_{2.5}$, respectively. As expected, ozone concentrations exhibit a distinct summer peak driven by increased solar radiation. The strong correlation between PM_{10} and $PM_{2.5}$ arises from their definitions, as $PM_{2.5}$ (particles with diameters below 2.5 µm) is a subset of PM_{10}.

2.2.2. Meteorological data

This dataset includes hourly measurements of relative humidity R_H, irradiance S, precipitation P, temperature T, wind direction W_d and wind speed W_s measured at 8 meteorological stations across Madrid. Fig. 1 shows the location of 5 of these stations in the central area of the city as red markers. The full list of meteorological stations, including their ID, name, location (latitude and longitude), and address, are shown in Table A.2 in Appendix A (see also Fig. A.1 for a complete map).

2.2.3. Road traffic data

This dataset includes traffic volume data expressed as a number of passing cars every 15 min, measured at each of the several thousand sensors scattered across the city of Madrid. Fig. 1 shows the road traffic intensity measuring locations (green dots) within 500 meters around each pollutant station (dashed line). The hourly vehicle count is then computed by taking the average over all measuring points within this vicinity area around each pollutant station and all 4 15 min periods.

This information is particularly relevant given that traffic-related emissions account for approximately 78% of local source emissions for NO_2 (59% of the total emissions) in Madrid (Borge et al., 2014). This underscores the importance of accurately capturing traffic intensity near air quality monitoring stations. While traffic volume is commonly used as a proxy for traffic-related emissions, it has notable limitations. Specifically, it does not account for factors such as vehicle speed, fuel type, or fleet composition, all of which significantly influence actual emission levels. As a result, traffic volume provides only a partial representation of traffic emissions. This limitation should be taken into account when interpreting the model results.

2.2.4. Per capita electricity consumption

To account for additional sources of local emissions beyond traffic, we incorporated three new predictors into the model: hourly power consumption in the *industrial*, *residential*, and *services* sectors. These variables serve as meaningful proxies for localized anthropogenic activity and are particularly relevant for modeling the pollutants, which originate from a wide range of emission sources.

The power consumption data were sourced from publicly available energy usage reports, which provide sectoral breakdowns of electricity demand in the Madrid region. These predictors help capture temporal variations in local activity patterns, such as increased residential heating during winter months or industrial operations during working hours, that can significantly influence particulate matter concentrations.

2.2.5. Additional features

To account for daily, weekly, and seasonal effects, the final set of predictors has been extended with the hour of the day, the day of the week, and the day of the year. Temporal features such as hour of the day, day of the week, and day of the year were incorporated into the model to account for the well-documented periodicity in air pollutant behavior. These variables help capture both diurnal patterns (e.g., morning and evening rush-hour peaks in NO_2) and weekly fluctuations (e.g., reduced traffic emissions on weekends). Additionally, the inclusion of the day of the year enables the model to account for seasonal changes in atmospheric conditions, photochemical activity, and human activity patterns that influence pollutant levels. To further illustrate the relevance of these temporal predictors, we include additional figures in Appendix D that visualize the variation of key variables (NO_2, traffic volume, temperature, and residential power consumption) across different time scales—hourly, weekly, and monthly. These plots highlight clear periodic patterns and support the inclusion of temporal features in the model. They demonstrate how pollutant levels and their main drivers change over time, reinforcing the temporal structure embedded in the modeling strategy.

To account for non-local effects on the levels of suspended particulate matter ($PM_{2.5}$ and PM_{10}), we also included an additional feature consisting of the background concentration of each pollutant as measured by a monitoring station in the rural outskirts of Madrid.

2.3. Modeling strategy

After pre-processing all data to detect and remove outliers and corrupted data points using a percentile-based clipping method (0.5th–99.5th percentiles), the working dataset has been built by combining the pollutant, meteorological, and road traffic intensity datasets using their hourly timestamp. For each pollutant species and measuring site, each observation contains an hourly concentration value, an average road traffic intensity (estimated with the data points within a 500-meter radius), and the meteorological predictors retrieved from the closest meteorological station. The working dataset contains approximately 10^5 data points (rows) for each pollutant, though the exact number varies depending on data completeness.

Data retrieval, analysis, pre-processing, modeling, post-processing, and visualization have been carried out using the programming language R 4.3.3. The library *tidymodels* (1.2.0), a comprehensive framework for streamlined modeling workflows (Kuhn and Silge, 2022), has

been used across the entire Machine Learning workflow. First, data is split into training and test sets. This separation ensures an unbiased evaluation of the model's performance, with the test data held separate from the modeling process. Given that our dataset spans from January 2022 to December 2023 (24 months), we generate a training dataset by randomly selecting 23 months of data and reserving one month for the test dataset.

Each model predictive performance has been measured using three different metrics: the coefficient of determination (R^2), the root mean squared error (RMSE), and the Bias. The R^2 defined in Eq. (1) quantifies the proportion of variance in the dependent variable explained by the independent variables, providing insight into the model's explanatory capacity. A higher R^2 value closer to one indicates a better fit of the model to the data. The RMSE defined in Eq. (2) measures the average magnitude of errors between predicted and actual values, offering a tangible measure of prediction accuracy, with a lower RMSE indicating better performance. Finally, the Bias – or mean error – defined in Eq. (3) indicates the model's tendency to consistently overestimate or underestimate values, with zero Bias being ideal.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}} \tag{2}$$

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i) \tag{3}$$

Here, n is the number of observations, y_i is the observation value i, f_i is the predicted value for observation i, and $\bar{y}$ is the mean of the observed values.

2.3.1. Feature engineering

To enhance data quality and improve model performance, we applied various feature engineering techniques. First, we reduced the significant values of skewness (SK) exhibited by the concentration of NO_2 and O_3 with values characterized by very asymmetrical distributions around the mean. Since values of these two pollutant concentrations are always above zero for all observations across all stations, we used the logarithmic transformation. Secondly, categorical predictors for the day of the week, the day of the year, and public holidays in Madrid were transformed using a one-hot encoding, and missing values were imputed using the median. Additionally, all numerical predictors were normalized by dividing by its mean and subtracting the standard deviation.

Furthermore, to capture the cyclic nature of certain predictors, we applied trigonometric transformations to features that would otherwise be treated as linear. Specifically, the hour of the day $h \in 0 \ldots 23$ was accounted via two new predictors, h_1 and h_2, defined as:

$$\begin{aligned} h_1 &= \sin\left(2\pi\frac{h}{24}\right) \\ h_2 &= \cos\left(2\pi\frac{h}{24}\right) \end{aligned} \tag{4}$$

Analogously, wind speed (W_s) and direction (W_d) were transformed using trigonometric functions to represent the wind vector in the north–south (W_y) and west–east (W_x) directions, enhancing the model's ability to capture directional wind effects.

$$\begin{aligned} W_y &= W_s \sin\left(\pi\frac{W_d}{180}\right) \\ W_x &= W_s \cos\left(\pi\frac{W_d}{180}\right) \end{aligned} \tag{5}$$

To summarize, Table 1 presents all predictors used in the model, categorized into traffic data, meteorological data, time-based predictors, and background monitoring data.

2.3.2. Learner

In the context of machine learning, a learner refers to an algorithm or computational model that is designed to learn patterns, relationships, and representations from data. In this study, we employed a diverse set of learners to predict pollutant levels and evaluate their performance, aiming to assess the impact of different learners on model performance. The selected learners included LightGBM (LGBM), random forest (RF), neural network (NN), Cubist, and XGBoost.

2.4. Cross-validation and hyper-parameter tuning

To robustly estimate the predictive capabilities of the different learners, data has been resampled into ten *folds*, each further divided into analysis and assessment sets. Each learner was then trained on the analysis set and evaluated on the independent assessment set for each resample.

Additionally, the most relevant hyper-parameters in each learner have been tuned to maximize the predictive performance during the model training. Thus, for instance, the number of trees in the ensemble and its depth along with the learning rate and the number of features that are randomly selected at each split (*mtry*) have been subjected to random search to decide upon the best combination of these hyper-parameters for the LGBM learner.

The benchmark for all five already-tuned learners is shown in Fig. 3. The vertical and horizontal axis corresponds to the pollutant measurement station (see Fig. A.1 for locations) and the learner, respectively. Tile color and the text label indicate, for each of the four pollutant species and location, the value of the correlation coefficient defined in Eq. (1). The performance metrics reported in Fig. 3 represent the average results obtained from a 10-fold cross-validation procedure, ensuring robustness and mitigating the influence of specific data partitions on model evaluation. Results suggest that LGBM offers the best overall performance across pollutants and measurement sites. R^2 scores for this learner across all available locations range between 0.79 to 0.86, 0.82 to 0.86, 0.73 to 0.90 and 0.57 to 0.72 for NO_2, O_3, PM_{10} and $PM_{2.5}$ respectively. The lower performance found for $PM_{2.5}$ is likely due to this pollutant characteristic of high variability and complex sources that are not accounted for through the current set of predictors. In addition, when comparing the R^2 values across locations, LGBM is found to offer very consistent predictive performance with relatively narrow score variance. Regarding the other learners, expect Neural Networks that clearly fall behind in terms of accuracy, Cubist, Random Forests, and XGBoost are found to perform almost as well as the LGBM. Analogous results for RMSE and Bias are shown in Figs. B.1 and B.2 in Appendix B. The overall conclusions extracted from these two additional metrics are very similar: LGBM is the best performer closely followed by the Cubist, Random Forests, and XGBoost.

LGBM was selected as the modeling algorithm for predicting pollutant concentration levels due to its efficiency and strong predictive performance. Hyperparameter optimization was performed using randomized search combined with cross-validation to ensure robust model tuning. The final LGBM configuration typically included 1,669 trees, a learning rate of 0.021, a maximum tree depth of 16, and 10 features evaluated at each split (mtry). This configuration was chosen to strike an effective balance between model accuracy and interpretability.

3. Performance analysis

After preprocessing and hyperparameter tuning, the predictive performance of the LGBM model was evaluated on unseen test data. This section is structured into five parts: (i) model accuracy across pollutants and locations, (ii) impact of incorporating background particulate matter, (iii) uncertainty estimation, (iv) feature importance analysis, and (v) comparison between the present model and other data-driven approaches for predicting the air quality in Madrid.

Table 1
List of predictors used in the model.

Category	Predictors	Symbol	Unit
Traffic Data	Traffic intensity	N_c	Cars per hour
Meteorological Data	Temperature	T	°C
	Relative humidity (%)	R_H	–
	Precipitation	P	mm
	Solar irradiance	S	W m^{-2}
	Wind vector	W_x, W_y	m s^{-1}
Time-Based Predictors	Hour of the day	h_1, h_2	–
	Day of the week	d_w	–
	Day of the year	d_y	–
	Madrid public holidays	h_d	–
Background Monitoring Data	$PM_{2.5}$	$\widehat{PM}_{2.5}$	μg m^{-3}
	PM_{10}	$\widehat{PM}_{10}$	μg m^{-3}
Power Consumption Data	Industrial sector electricity usage	PC_I	kW h
	Residential sector electricity usage	PC_R	kW h
	Services sector electricity usage	PC_S	kW h

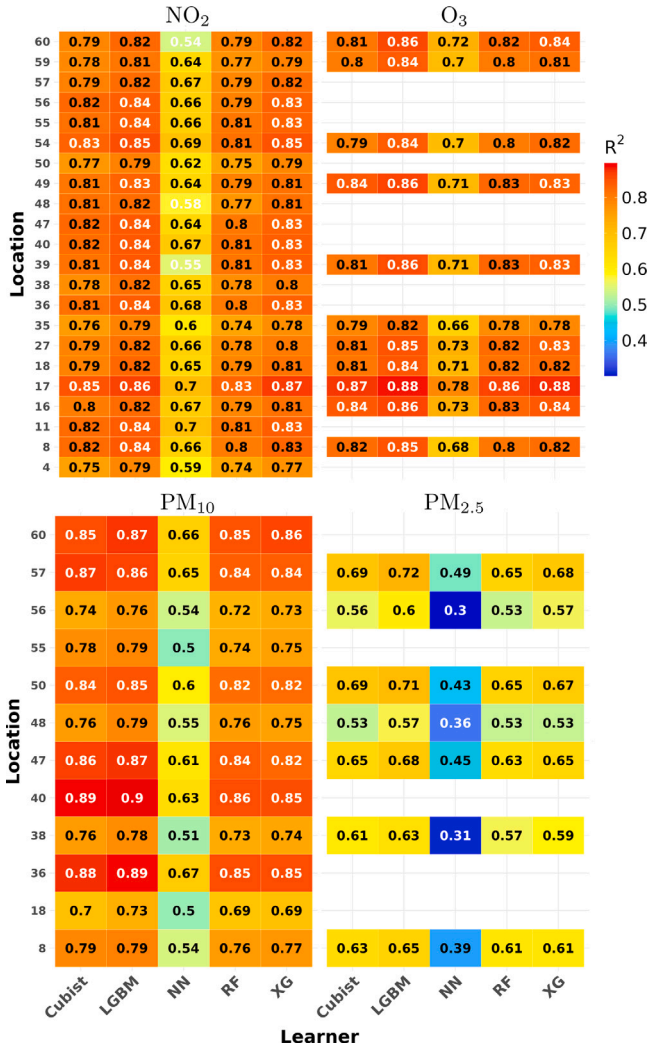

Fig. 3. Values of R^2 (color and label) for the five learners considered in this work for each pollutant measurement site and pollutant species. Missing data indicated a lack of data for that given pollutant and location.

3.1. Model accuracy across locations and pollutants

Fig. 4 presents the evaluation metrics for the test dataset. The top-left panel displays the values of R^2_{te}, where the subscript *te* denotes the test data. Pollutant species are represented on the horizontal axis, while measuring sites are on the vertical axis. For NO_2, the model demonstrates strong predictive performance, with R^2_{te} values ranging from 0.77 to 0.86, indicating accurate concentration estimates across Madrid. A similar performance is observed for ozone, despite being available at fewer sites, with R^2_{te} values between 0.8 and 0.85. Predictions for PM_{10} exhibit generally strong performance, with R^2_{te} values ranging from 0.81 to 0.95 across most locations, except for location 18, where the score drops to 0.68. Finally, $PM_{2.5}$ remains the most challenging pollutant to predict, with R^2_{te} values ranging from 0.63 to 0.82.

We observed variability in PM_{10} and $PM_{2.5}$ prediction performance across different monitoring sites. Specifically, the lowest performance for PM_{10} occurred at location 18 (Farolillo), while $PM_{2.5}$ performance was weakest at locations 48 (Castellana) and 56 (Plaza Elíptica). Although our dataset does not include detailed information on local emission sources or micro-environmental factors at each site, these differences may be attributed to complex urban conditions such as proximity to unaccounted emission sources (e.g., construction zones), building geometry effects on air dispersion, or site-specific meteorological patterns. Moreover, the model's poor performance at location 48 (Castellana) for $PM_{2.5}$ may also be partially explained by data quality issues. This site exhibited the highest proportion of missing values in $PM_{2.5}$ measurements among all locations, which may have affected both the training and evaluation processes.

The top-right panel presents the $RMSE_{te}$ values, providing an analogous evaluation of the correlation coefficient. As observed with R^2, the RMSE exhibits relatively small variance across locations, indicating strong model robustness throughout Madrid. Unlike R^2, RMSE is not normalized, meaning its magnitude is directly influenced by pollutant concentration levels. Consequently, the lowest RMSE values are found for $PM_{2.5}$, which has the lowest average concentration (see Fig. 2). In contrast, ozone, the pollutant with the highest average concentration, also exhibits the largest RMSE values. Overall, $RMSE_{te}$ ranges between 3.2 and 13 μg m^{-3}.

Also, a non-normalized metric, the BIAS magnitude, shown in the bottom left panel, reflects the differences in average concentration levels between pollutants. The mean deviation between observed and predicted values ranges from 0.52 to 3.5, 1.3 to 3.2, −2 to 0.8, and −0.3 to 0.5 μg m^{-3} for NO_2, O_3, PM_{10}, and $PM_{2.5}$, respectively. While the BIAS distribution for nitrogen dioxide and ozone is positively skewed across all locations, the model tends to slightly overpredict particulate matter concentrations at certain sites.

Finally, the bottom right panel in Fig. 4 shows the ratio of the correlation coefficients between the training and test datasets. This metric ideally approaches unity, indicating that the models are not substantially affected by overfitting. While this ratio is consistently equal to or greater than 0.77 for NO_2, O_3, and PM_{10}, it is a bit lower for $PM_{2.5}$, reinforcing the finding that this pollutant is the most challenging to model accurately.

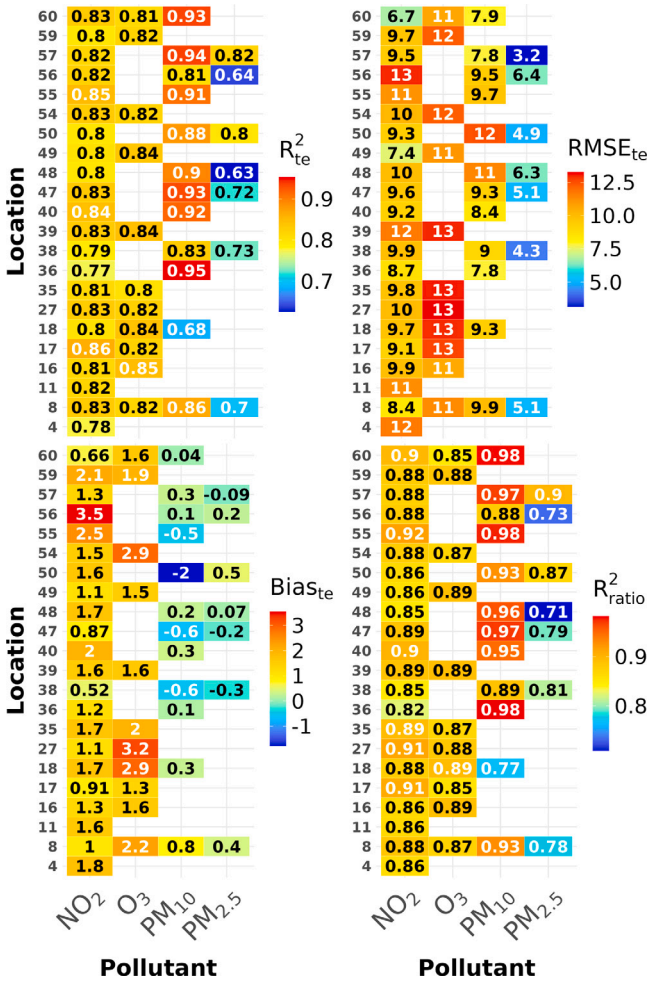

Fig. 4. Metric values for different pollutants at different locations.

3.2. Effect of including background PM

To explore the influence of regional pollution, we compared two model versions: with and without background $PM_{2.5}$ and PM_{10} values as predictors. Background concentrations of $PM_{2.5}$ and PM_{10} were incorporated as hourly values obtained from official air quality monitoring stations in Madrid, ensuring temporal consistency with the other predictor variables.

The ratio of performance metrics for the cases with and without the rural background levels of particulate matter as a predictor is shown in Fig. 5 for $PM_{2.5}$ (top panel) and PM_{10} (bottom panel). Overall, including the background far-field particulate matter levels improves the prediction of PM concentrations within the city of Madrid, as evidenced by R^2 values greater than unity, with several locations exhibiting ratios above 1.3. Similarly, incorporating the background field leads to lower RMSE values, with reductions in the ratio ranging from 0.53 to 0.98.

The variability observed in the RMSE ratio can be attributed to differences in the contribution of background concentrations across locations. Regional background levels play a dominant role in some areas, significantly enhancing model performance when included. In contrast, other locations may be more influenced by unaccounted-for local sources, such as construction activities, industrial emissions, or localized resuspension of road dust, where background monitoring data contribute less.

3.3. Model uncertainty

To enhance the robustness and interpretability of the model predictions, we further performed an uncertainty analysis using quantile regression based on the LightGBM framework. This analysis provides prediction intervals for pollutant concentrations. Details of the methodology and results of this uncertainty assessment are provided in Appendix C.

3.4. Feature importance analysis

Beyond their predictive capabilities, the models developed for each pollutant and location also provide valuable insights into the relative contribution of different predictors to local air pollution levels. This is assessed through Feature Importance (FI) analysis, which quantifies the impact of each variable in explaining the observed pollutant concentrations.

Fig. 6 presents the FI results for NO_2, O_3, $PM_{2.5}$, and PM_{10} across different monitoring sites. The color scale represents the relative FI weights, with predictors listed on the horizontal axis and measurement sites on the vertical axis. To focus on the most influential factors, only the top 13 predictors are included: the number of cars (N_c), day of the year (d_y), hour of the day (h_1), solar irradiance (S), background concentrations of particulate matter ($\widehat{PM}_{10}$ and $\widehat{PM}_{2.5}$), air relative humidity (R_H), air temperature (T), power consumptions in three different sectors (PC_I, PC_S, and PC_R) and the East-West and North-South wind components (W_x and W_y, respectively).

The FI analysis suggests that the relative importance of predictors for NO_2 remains fairly consistent across all monitoring locations, with relatively narrow distributions of weight values. Among all variables, wind vectors play the most significant role in determining NO_2 concentrations in most urban areas, likely due to their influence on pollutant dispersion and accumulation. Additionally, the number of cars and the day of the year are key factors, reflecting the dominant contribution of road traffic emissions and their seasonal variation. Following in importance, solar irradiance emerges as a relevant predictor, highlighting its role in driving photochemical reactions that influence nitrogen dioxide levels.

The role of photochemistry in shaping the nitrogen oxides-ozone balance is evident in the FI results for O_3. Unlike NO_2, a primary pollutant largely emitted by traffic, ozone is a secondary pollutant, primarily formed through photochemical reactions involving nitrogen oxides and volatile organic compounds in the presence of sunlight. Day of the year is among the dominant predictors. However, the relative importance of other variables varies by location. In most monitoring sites, wind components (W_x, W_y) and temperature (T) are more influential than solar irradiance (S). Nevertheless, at certain locations (8, 49, and 54), solar irradiance plays a more prominent role, reflecting localized differences in photochemical activity or atmospheric conditions. As anticipated, the model shows that traffic density has minimal direct influence on O_3 concentrations, which is consistent with the fact that ozone levels are primarily governed by regional and large-scale photochemical processes, with local sources playing a secondary role.

For smaller particles ($PM_{2.5}$), the results indicate that background pollution levels, day of the year, and temperature are the most influential predictors, highlighting the combined effect of regional transport, seasonal variations, and meteorological conditions on particulate matter concentrations. However, the FI results also exhibit notable variability across different locations, suggesting that local sources and site-specific factors play a role in shaping $PM_{2.5}$ levels.

In contrast, the FI results for PM_{10} reveal a markedly different pattern. Here, background PM_{10} concentrations emerge as the overwhelmingly dominant predictor, with weight values often an order of magnitude higher than those of any other variable. Unlike the other three pollutants, for which FI weights exhibit considerable spatial variability, the relative importance of predictors for PM_{10} remains remarkably consistent across all locations. This suggests that

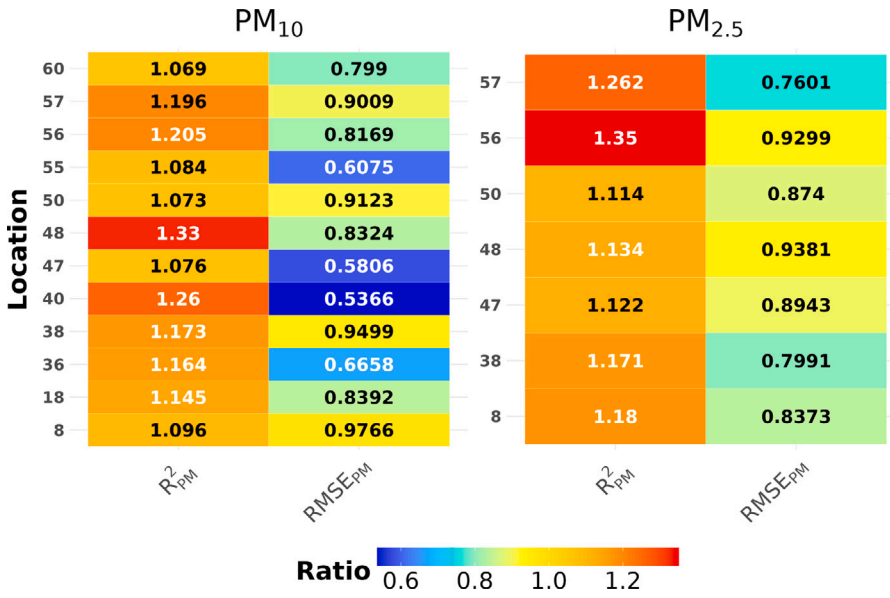

Fig. 5. The ratio of R^2 and RMSE between the cases with and without the background particulate matter concentration as a predictor for both $PM_{2.5}$ and PM_{10} at each location.

particulate pollution in urban environments is largely driven by regional background levels rather than local sources or meteorological conditions.

In addition to individual feature importance, we analyzed interaction effects between predictors using the iml package in R. Results and visualizations are provided in Appendix C.

To evaluate the performance of the methodology presented in this study, the model's accuracy has been assessed by comparing its RMSE and BIAS values for NO_2, O_3, PM_{10}, and $PM_{2.5}$ against those of other modeling approaches. However, this comparison should be interpreted with caution due to the differences in methodology, scope, and framework among the various models.

3.5. Comparison between the present model and other data-driven approaches

The comparison between our model and other forecasting tools, such as CAMS and SOCAIRE, was conducted using data from the same air quality monitoring stations employed in our study. For CAMS, the first 48 h of their four-day forecast were considered to ensure consistency with the forecast horizon of the SOCAIRE model. Unlike these systems, our model does not rely on prior pollutant concentration data and can provide hourly predictions for any given time, based solely on the corresponding predictor values (e.g., traffic, meteorology, and power consumption). Unlike CAMS, which provides concurrent forecasts of both meteorological conditions and air pollutant concentrations, our model relies on the availability of input predictor values at the target prediction time. Therefore, the model does not forecast into the future in a temporal sense but rather estimates pollutant concentrations for a specified hour, given the corresponding input conditions.

Despite these methodological differences, the model demonstrates excellent performance, particularly in predicting NO_2 and O_3, where it outperforms many alternative approaches. For particulate matter (PM_{10} and $PM_{2.5}$), the results are more modest but remain competitive.

Table 2 displays the RMSE and bias values for the proposed methodology, comparing it against five alternative models: CAMS (Copernicus Atmosphere Monitoring Service), LR (Linear Regression), NNED (Neural Network Encoder-Decoder), Persistence, and SOCAIRE, all of which were implemented by de Medrano et al. (2021). CAMS (Peuch et al., 2022) acts as a significant baseline, offering four-day hourly pollution forecasts across Europe through synoptic-scale modeling. It combines global and regional numerical weather predictions from ECMWF (Marécal et al., 2015), alongside forecasts of both natural and anthropogenic chemical production. While CAMS provides extensive coverage, a localized model would ideally achieve improved accuracy by capturing finer spatial and temporal variations. LR is a commonly used, simple regression method that performs well in many contexts, although its linear assumptions limit its flexibility. It serves as a baseline model, establishing a minimum performance standard, similar to persistence. de Medrano et al. (2021) employed a multi-output approach, utilizing separate models for each station to predict multiple future time steps based on past values. The NNED model, a specialized convolutional neural network encoder–decoder, predicts pollutant concentrations across all stations using imputed data, numerical weather predictions (NWP), numerical pollution predictions (NPP), and historical pollution levels. It effectively captures non-linear relationships and feature interactions but struggles to account for irregularities in non-cyclical anthropogenic factors, such as traffic variations. The persistence model, a simple benchmark, assumes pollutant concentrations remain unchanged from the previous time step. Its performance serves as a reference point for evaluating the improvements offered by more advanced models. de Medrano et al. (2021) also introduced an enhanced version of persistence that incorporates daily, weekly, and annual cyclical patterns to better capture temporal fluctuations. SOCAIRE, an advanced air quality forecasting and monitoring tool currently in use in Madrid, is based on a Bayesian and spatiotemporal ensemble of neural and statistical nested models. It integrates both endogenous and exogenous factors, including past pollutant concentrations, human activity, and numerical weather and pollution predictions, to model and predict air quality dynamics. SOCAIRE provides 48-hour probabilistic forecasts for key pollutants, estimating full probability distributions for compound events.

The proposed modeling framework offers significant advantages in terms of cost-efficiency and accessibility. It is built using LightGBM, a highly efficient gradient boosting algorithm that enables rapid training and prediction, with reported speeds 2–10 times faster than other popular learners. This makes the model practical for real-time applications and suitable for deployment on standard computing infrastructure. Moreover, all input features – including air quality, meteorological data, traffic intensity, and power consumption – are obtained from open-access sources, ensuring that the model can be easily adapted and transferred to other urban environments with similar data availability. Unlike traditional forecasting systems such as SOCAIRE, which rely on

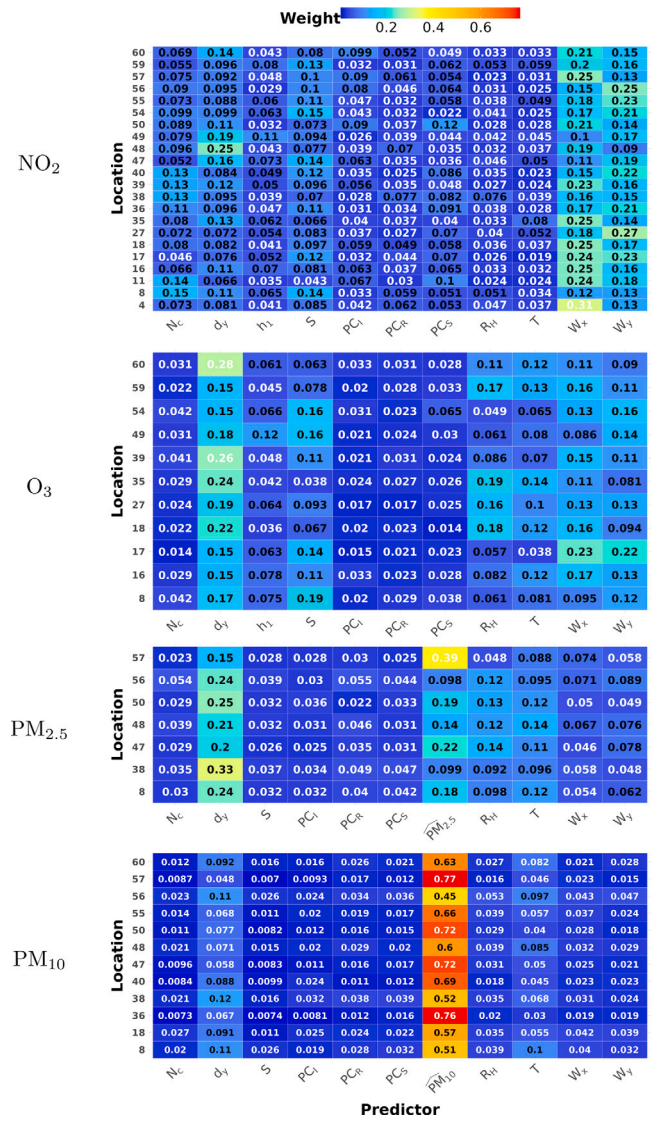

Fig. 6. Feature importance analysis for different pollutant concentration predictions at different locations.

historical pollutant trends, our model can predict pollutant concentrations for any given time based solely on current input features. This enhances its operational flexibility and makes it particularly useful in settings where historical data may be sparse or unreliable.

4. Sensitivity analysis of traffic reduction on air quality

To demonstrate potential applications of the present model, we estimate the impact of changes in road traffic intensity on ambient NO_2 concentrations. In urban environments, fossil fuel-based transportation is a primary source of nitrogen dioxide emissions, primarily due to combustion processes in vehicles. Understanding how NO_2 levels respond to changes in traffic volume is crucial for designing effective air quality management strategies.

Table 3 presents the estimated percentage change in average NO_2 concentrations resulting from a hypothetical 10% reduction in the number of cars across the city. The results are shown for six selected days of the year and include values aggregated across all monitoring locations as well as for two specific sites: location 35, situated in the city center, and location 57, representing a suburban area.

The findings indicate that the reduction in NO_2 levels is more pronounced in densely populated urban areas (location 35) compared to suburban settings (location 57). The observed decrease in NO_2 concentrations varies between 8.7% and 1.6%, depending on the day of the year. This variability may be attributed to several factors, including seasonal fluctuations in photochemical activity, which influence the nitrogen oxides and ozone equilibrium, as well as variations in traffic intensity throughout the year.

This type of analysis offers valuable insights for policymakers and urban planners by quantifying the expected air quality benefits of traffic reduction measures. The results highlight that NO_2 concentrations respond differently based on both spatial (urban vs. suburban) and temporal (seasonal) variations, emphasizing the need for location-specific air pollution mitigation strategies. Further exploration of these dynamics could help refine traffic management policies, optimize emissions control measures, and improve public health outcomes in urban areas.

It is important to note that, while our model shows strong predictive capabilities, it is fundamentally data-driven and not physically based. As such, the sensitivity analysis presented should be interpreted as a reflection of the statistical relationships learned by the model rather than as physically causal mechanisms. While this analysis provides a preliminary indication of potential air quality improvements under reduced traffic scenarios, it is important to interpret the results cautiously. Further investigation using more comprehensive scenario analysis and multiple levels of reduction is needed before drawing robust conclusions regarding policy implications.

Table 2
Comparison of metric values for NO_2, O_3, $PM_{2.5}$, and PM_{10} predictions.

	NO_2		O_3	
	RMSE	Bias	RMSE	Bias
CAMS	23.5 (9.1)	12.3 (9.8)	19.1 (5.0)	-3.2 (6.2)
LR	20.7 (6.3)	-0.1 (1.8)	20.5 (3.9)	-0.8 (1.1)
NNED	16.5 (4.5)	-9.2 (3.2)	16.8 (1.4)	2.4 (3.3)
Persistence	26.4 (9.3)	-1.4 (3.5)	27.4 (4.8)	0.5 (16)
SOCAIRE	14.9 (4.8)	-0.2 (0.8)	15.8 (2.8)	1.6 (1.0)
Present work	9.83 (3.13)	0.22 (0.76)	11.89 (1.35)	-0.3 (0.9)
	PM_{10}		$PM_{2.5}$	
	RMSE	Bias	RMSE	Bias
CAMS	13.9 (3.7)	6.3 (3.7)	6.5 (1.3)	1.1 (2.1)
LR	13.4 (2.9)	0.3 (0.7)	7.2 (1.3)	0.2 (0.5)
NNED	14.2 (1.3)	2.4 (1.8)	5.6 (0.8)	-1.4 (0.4)
Persistence	15.5 (3.1)	-0.6 (3.5)	7.8 (1.4)	-0.3 (1.4)
SOCAIRE	10.6 (2.5)	0.3 (0.7)	5.4 (1.0)	-0.1 (0.6)
Present work	10.14 (2.32)	0.6 (1.4)	4.76 (1.61)	0.11 (0.07)

Table 3
Percentage change in NO_2 concentrations due to a 10% reduction in traffic.

Day of the year	% Change (Citywide)	% change at 35	% change at 57
1	−1.53	−3.11	−1.38
60	−5.75	−3.64	−2.27
120	−3.53	−2.22	−1.99
180	−8.72	−5.7	−3.12
240	−2.89	−3.98	−2.22
300	−5.24	−3.33	−3.12

5. Conclusions

This study presents a parsimonious, data-driven model for predicting air pollutant concentrations in Madrid using only open data on meteorological conditions, power consumption data, and road traffic intensity. Unlike many existing Machine Learning (ML) models that rely on complex architectures and heterogeneous data sources – including historical pollution records – this approach demonstrates that access to highly relevant predictor data can yield competitive or superior performance with significantly lower computational demands. The model

is designed to estimate pollutant concentrations at a specific time point based solely on the corresponding values of input predictors (e.g., meteorological variables, traffic intensity, power consumption), rather than using historical pollutant data or time-lagged features. As such, it serves as a flexible, scenario-based estimation tool rather than a traditional time series forecasting model.

The model achieves strong predictive accuracy, particularly for NO_2 and O_3, where it outperforms or matches more complex solutions. For particulate matter (PM_{10} and $PM_{2.5}$), results remain competitive but exhibit greater variability across locations. The study highlights the crucial role of localized traffic data in improving prediction accuracy, particularly in areas where road transport is the dominant source of emissions.

Beyond its predictive capabilities, the model also enables impact assessments of traffic-related pollution control measures. A sensitivity analysis estimating the effects of a 10% reduction in road traffic suggests that NO_2 concentration reductions are more pronounced in high-density urban areas compared to suburban locations. These findings reinforce the spatial dependence of traffic emissions and their role in shaping local air quality.

Despite being limited to predictions at measurement locations, this work lays the foundation for future efforts to develop geospatially continuous air quality models. Potential extensions include integrating spatial interpolation techniques such as Kriging or Stochastic Partial Differential Equation (SPDE) models, which could enable city-wide pollutant concentration estimates at unmonitored locations. Based on the promising results of our model, particularly its high accuracy, simplicity, and reliance on accessible open data, it shows potential for future development in areas such as real-time air quality assessment, scenario analysis for urban planning, or as a complementary tool in early warning systems. However, we acknowledge that further validation and refinement are needed before such applications can be fully realized. As such, these future directions should be considered as informed possibilities rather than definitive outcomes of the present study.

Overall, this study demonstrates that open-data-driven, computationally efficient models can serve as valuable, accessible tools for air quality assessment and policy-making. By leveraging real-time meteorological and traffic data, this approach provides an interpretable and scalable framework that can support evidence-based decision-making in urban air quality management.

CRediT authorship contribution statement

Koorosh Kazemi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Anton Vernet:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Alexandre Fabregat:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Koorosh kazemi reports financial support was provided by Spain Ministry of Science and Innovation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by project TED2021-129348B-I00, funded by MCIN/AEI/10.13039/501100011033, Spain and by the European Union NextGenerationEU/PRTR.

Appendix A. Stations information

This appendix presents detailed information about the locations used for air quality and meteorological data collection in Madrid. The data includes the station name, address, coordinates (latitude and longitude), and identification number (ID).

Madrid, the capital of Spain with a population exceeding 3 million, is one of the most densely populated and motorized cities in Europe. Like many urban areas, it faces persistent air quality challenges, particularly from nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM_{10} and $PM_{2.5}$). Road traffic has been identified as the dominant source of local emissions, especially for NO_2. According to Borge et al. (2014), traffic sources account for approximately 78% of local NO_2 emissions in the Madrid region. This makes traffic a critical component when assessing urban air quality dynamics.

Annual average NO_2 concentrations in central districts frequently exceed the European Union's air quality limit of 40 $\mu g/m^3$, particularly near major roadways and intersections. PM_{10} levels typically range between 15–35 $\mu g/m^3$, while $PM_{2.5}$ levels fall within 5–20 $\mu g/m^3$, depending on seasonal conditions and proximity to emission sources. Ozone levels, on the other hand, tend to peak during spring and summer due to photochemical activity, often exceeding the recommended thresholds for human health protection.

In addition to air quality concerns, Madrid is equipped with a robust network of open-access data sources, including real-time air quality monitoring stations, meteorological stations, and traffic flow counters, which makes it an ideal candidate for the development and validation of data-driven air quality models. By incorporating traffic data, meteorological parameters, and sector-specific power consumption as proxies for local activity and emissions, the proposed modeling framework leverages openly available data to construct a scalable and replicable approach to pollutant prediction.

Table A.1 provides information on the air quality monitoring stations where various pollutant concentrations were measured. These stations are located throughout Madrid and were selected to ensure comprehensive coverage of the urban environment for accurate air quality modeling.

Table A.2 lists the meteorological stations used for collecting weather-related data such as temperature, humidity, wind speed, and wind direction. These stations are also distributed across Madrid and contribute essential input variables to the predictive modeling of pollutant concentration.

The spatial arrangement of the air quality monitoring and meteorological locations is visually represented in Fig. A.1.

Fig. A.2 illustrates the data completeness for each air quality monitoring and meteorological station, shown as the percentage of available hourly measurements for each predictor in relation to the number of observations over a two-year period. This figure presents two heatmaps: the top panel corresponds to pollutant monitoring stations, and the bottom panel represents meteorological stations. Each cell shows the completeness percentage (CP) for individual pollutants or meteorological variables at specific monitoring locations, with the color intensity indicating the level of data coverage. This figure serves as a comprehensive guide to understanding the specific variables monitored at individual sites within the study area. While meteorological properties are examined at designated locations, comprehensive reporting of all meteorological variables is restricted to just 8 specific locations.

Table A.1
Air quality monitoring stations in Madrid.

ID	Station	Long.	Lat.	Direction
4	Plaza de España	−3.712	40.423	Plaza de España
8	Escuelas Aguirre	−3.682	40.421	Entre C/ Alcalá y C/ O'Donell
11	Ramón y Cajal	−3.677	40.451	Avda. Ramón y Cajal esq. C/ Príncipe de Vergara
16	Arturo Soria	−3.639	40.440	C/ Arturo Soria esq. C/ Vizconde de los Asilos
17	Villaverde	−3.713	40.347	C/ Juan Peñalver
18	Farolillo	−3.731	40.395	C/ Farolillo - C/ Ervigio
27	Barajas Pueblo	−3.580	40.477	C/ Júpiter
35	Plaza del Carmen	−3.703	40.419	Plaza del Carmen esq. Tres Cruces
36	Moratalaz	−3.645	40.408	Avda. Moratalaz esq. Camino de los Vinateros
38	Cuatro Caminos	−3.707	40.446	Avda. Pablo Iglesias esq. C/ Marqués de Lema
39	Barrio del Pilar	−3.711	40.478	Avda. Betanzos esq. C/ Monforte de Lemos
40	Vallecas	−3.652	40.388	C/ Arroyo del Olivar esq. C/ Río Grande
47	Méndez Álvaro	−3.687	40.398	C/ Juan de Mariana/Plaza Amanecer Méndez Álvaro
48	Castellana	−3.690	40.439	C/ José Gutiérrez Abascal
49	Parque del Retiro	−3.683	40.414	Paseo Venezuela- Casa de Vacas
50	Plaza Castilla	−3.689	40.466	Plaza Castilla (Canal)
54	Ensanche de Vallecas	−3.612	40.373	Avda. La Gavia/Avda. Las Suertes
55	Urb. Embajada	−3.581	40.462	C/ Riaño (Barajas)
56	Plaza Elíptica	−3.719	40.385	Plaza Elíptica - Avda. Oporto
57	Sanchinarro	−3.661	40.494	C/ Princesa de Éboli esq. C/ María Tudor
59	Juan Carlos I	−3.616	40.461	Parque Juan Carlos I (frente oficinas mantenimiento)
60	Tres Olivos	−3.689	40.501	Plaza Tres Olivos

Table A.2
Meteorological stations in Madrid.

ID	Station	Long.	Lat.	Direction
24	Casa de Campo	−3.747	40.419	Casa de Campo (Terminal del Teleférico)
54	Ensanche de Vallecas	−3.612	40.373	Avda. La Gavia/Avda. Las Suertes
59	Juan Carlos I	−3.616	40.461	Parque Juan Carlos I (frente oficinas mantenimiento)
102	J.M.D. Moratalaz	−3.637	40.397	C/ Fuente Carantona
103	J.M.D. Villaverde	−3.711	40.366	C/ Arroyo Bueno
106	Centro Mpal. De Acústica	−3.739	40.442	Autovía M-30 Km. 21.700
107	J.M.D. Hortaleza	−3.656	40.463	Ctra. de Canillas
108	Peñagrande	−3.718	40.476	C.D.M. Peñagrande

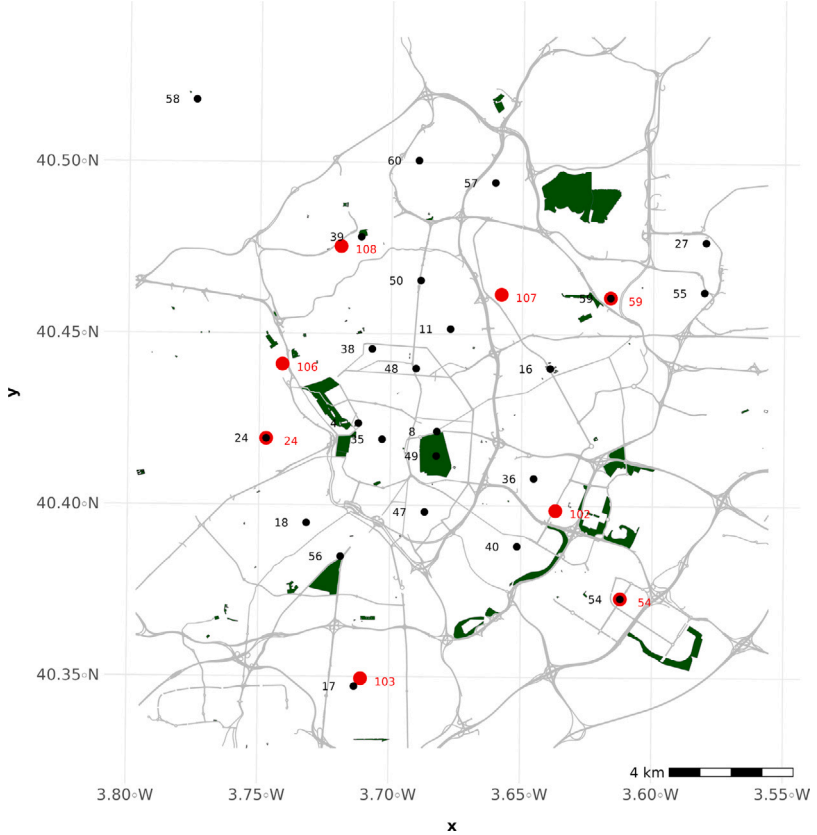

Fig. A.1. Spatial distribution: Air quality monitoring stations depicted by black points and meteorological stations indicated by red points on the map.

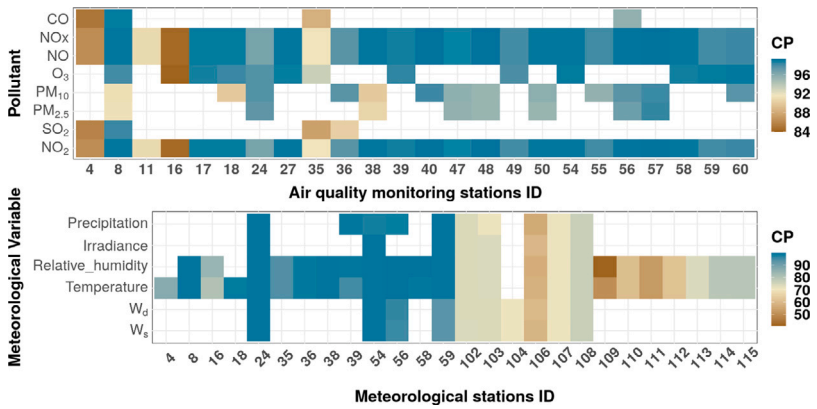

Fig. A.2. Pollutants and meteorological variables measured at each location.

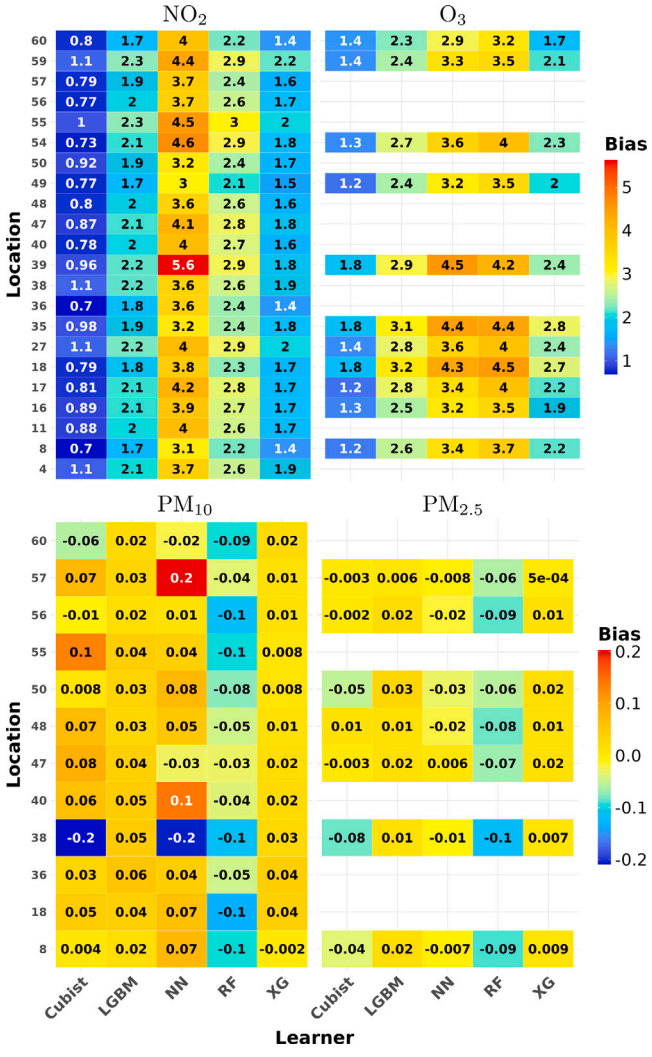

Fig. B.1. Bias values for different learners.

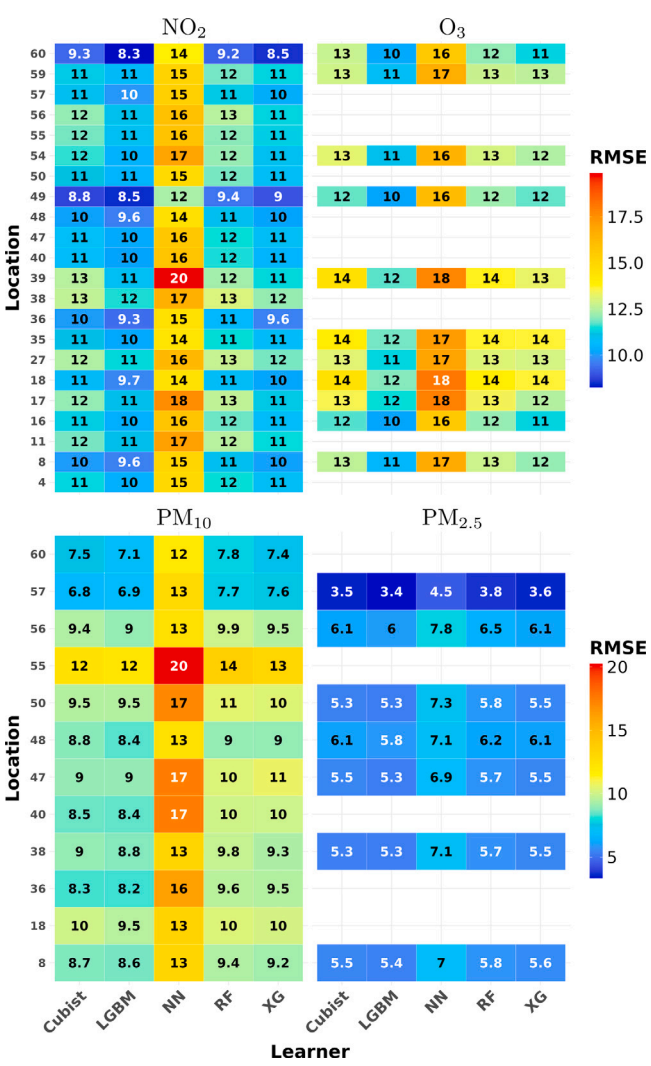

Fig. B.2. RMSE values for different learners.

Appendix B. Model prediction accuracy

This appendix provides a comparative analysis of different learners by evaluating their performance using Bias and RMSE metrics. It examines how each model performs in predicting pollutant concentrations and highlights variations in accuracy and systematic errors. Additionally, it presents a detailed assessment of the selected model's performance at specific air quality monitoring locations within the test dataset, offering insights into its reliability and predictive capability.

Figs. B.1 and B.2 provide a comparative analysis of different learners by evaluating their Bias and RMSE values, respectively. These figures illustrate how each model performs in terms of systematic error and overall predictive accuracy.

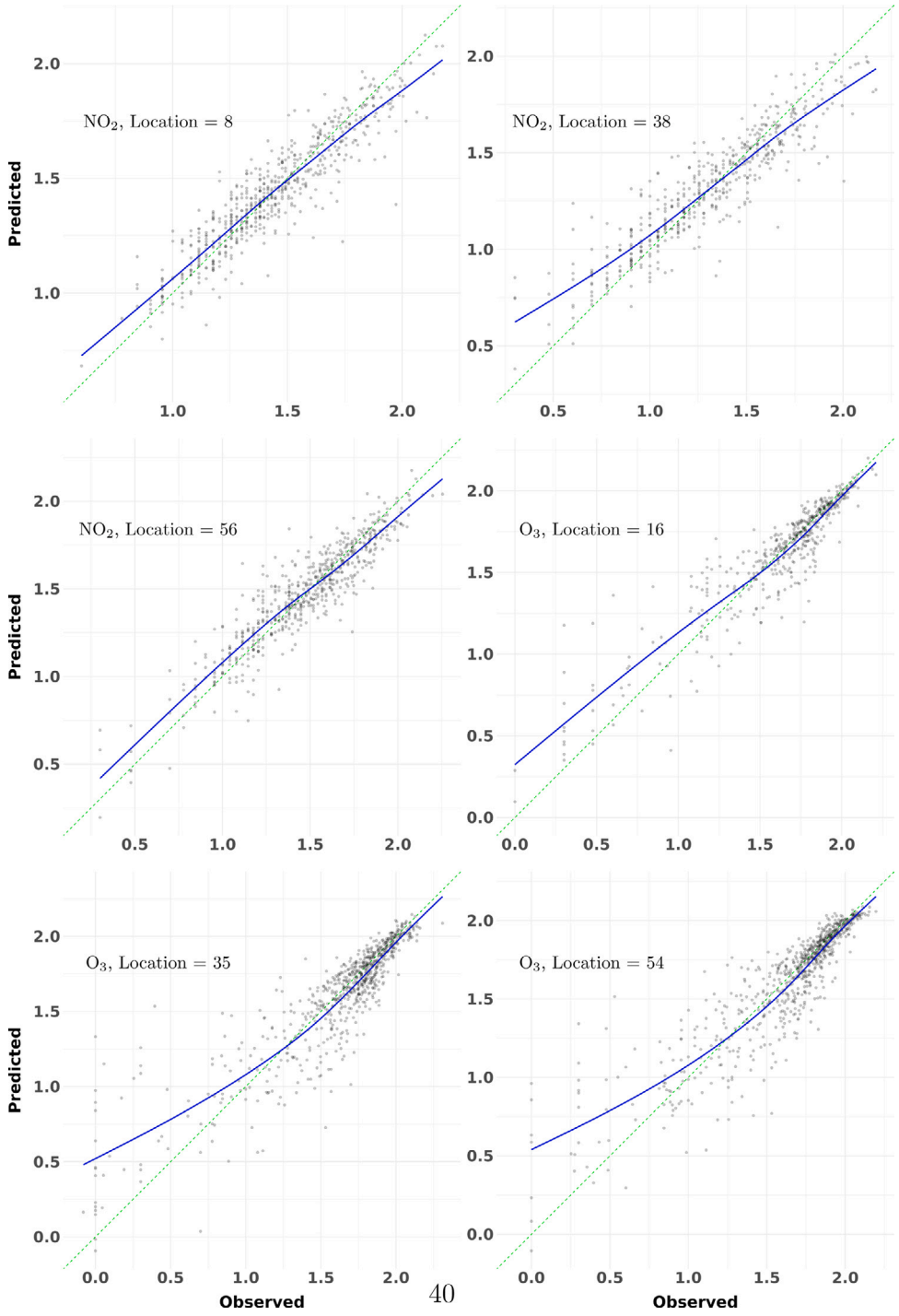

Fig. B.3. Predicted versus observed concentrations of NO_2 and O_3 in test dataset at selected locations. The blue line demonstrates the model's fitting line. The scales for NO_2 and O_3 are logarithmic.

Fig. B.3 presents scatter plots comparing observed and predicted values for NO_2 and O_3. The visual assessment of the model's accuracy is enhanced by two reference lines: a green dashed line representing the ideal scenario where predictions perfectly match observations, and a blue line indicating the actual fit achieved by our model. The closeness of the blue line to the ideal line reflects the model's capability to capture key patterns in the data and accurately predict pollutant levels, underscoring its effectiveness for air quality prediction tasks.

Fig. B.4 demonstrates the alignment between predicted and observed concentrations of $PM_{2.5}$ and PM_{10} in the test dataset at selected stations. The dashed green line denotes the optimal scenario where predictions precisely match observations, while the blue line represents the model's fitting line. The proximity of the blue line to the green dashed line highlights the model's efficacy in capturing underlying trends and reliably predicting pollutant concentrations.

Appendix C. Model reliability analysis

This appendix provides complementary analyses aimed at enhancing the reliability of the proposed model. While the main manuscript focuses on predictive performance and feature contributions, this section presents additional evaluations, including uncertainty quantification

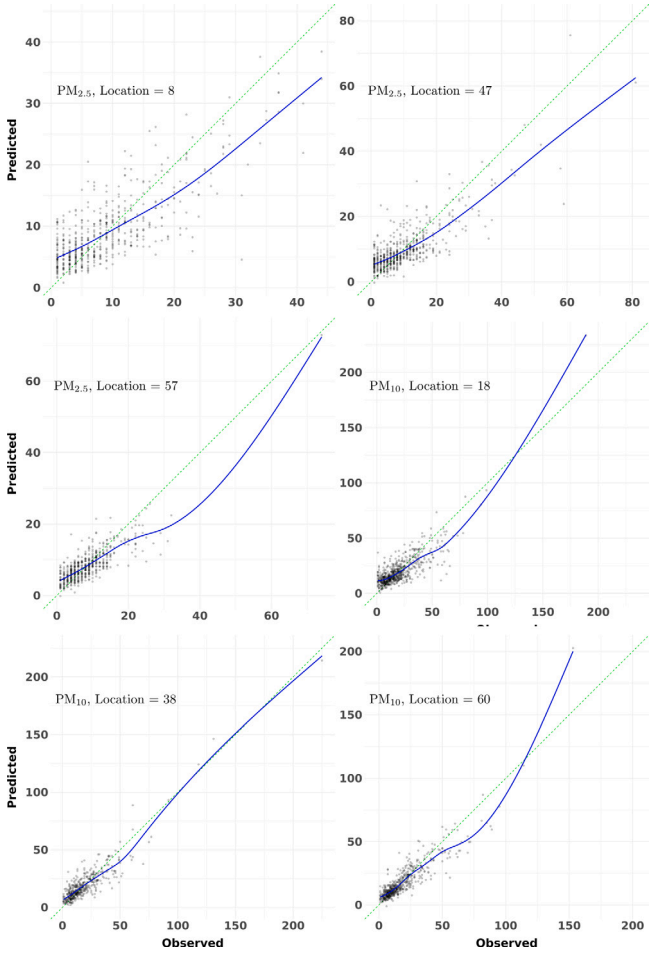

Fig. B.4. Predicted versus observed concentrations of $PM_{2.5}$ and PM_{10} in test dataset at selected locations. The blue line demonstrates the model's fitting line.

and predictor interaction analysis, to offer a more robust and comprehensive understanding of the model's behavior and decision-making mechanisms.

C.1. Uncertainty quantification via quantile regression

To assess the uncertainty in the predicted pollutant concentrations, we implemented quantile regression using the LightGBM framework. This approach goes beyond point estimates by modeling conditional quantiles of the response variable, offering a probabilistic view of prediction ranges.

Specifically, three separate models were trained to predict the 5th, 50th (median), and 95th percentiles of the pollutant concentrations. The 5th–95th percentile interval forms a 90% prediction interval, estimating the range within which the true pollutant concentration is expected to fall. These prediction intervals are crucial for informing decisions where uncertainty plays a key role, such as in public health response or air quality risk assessment.

Figs. C.1 and C.2 illustrate these prediction intervals for NO_2 (location 55), O_3 (location 59), $PM_{2.5}$ (location 8), and PM_{10} (location 8). We selected twelve representative time points – specifically, the 15th of each month of 2023 at 12:00 PM – to visualize seasonal variation and uncertainty across the year. Each point in the plots shows the median predicted concentration along with the shaded area indicating the 5th to 95th percentile range.

This visualization highlights not only how pollutant levels vary over the course of the year but also the degree of uncertainty in each estimate.

C.2. Interaction effects analysis

While feature importance helps identify the most influential predictors, it does not account for how variables may interact with each other. To capture these interaction dynamics, we performed an interaction analysis using the `iml` package in R, which provides model-agnostic interpretability tools compatible with tree-based models like LightGBM.

We computed interaction scores for each predictor, which quantify the extent to which a variable's effect depends on the values of other variables. Higher interaction scores indicate stronger dependency relationships between variables, revealing where predictor synergies significantly influence pollutant concentrations.

Figs. C.3 and C.4 display the interaction scores for all predictors used in modeling NO_2 (location 55), O_3 (location 59), $PM_{2.5}$ (location 8), and PM_{10} (location 8). These visualizations help uncover complex behaviors in the data, such as how traffic intensity might interact with meteorological features (e.g., temperature or wind speed) in shaping pollution levels.

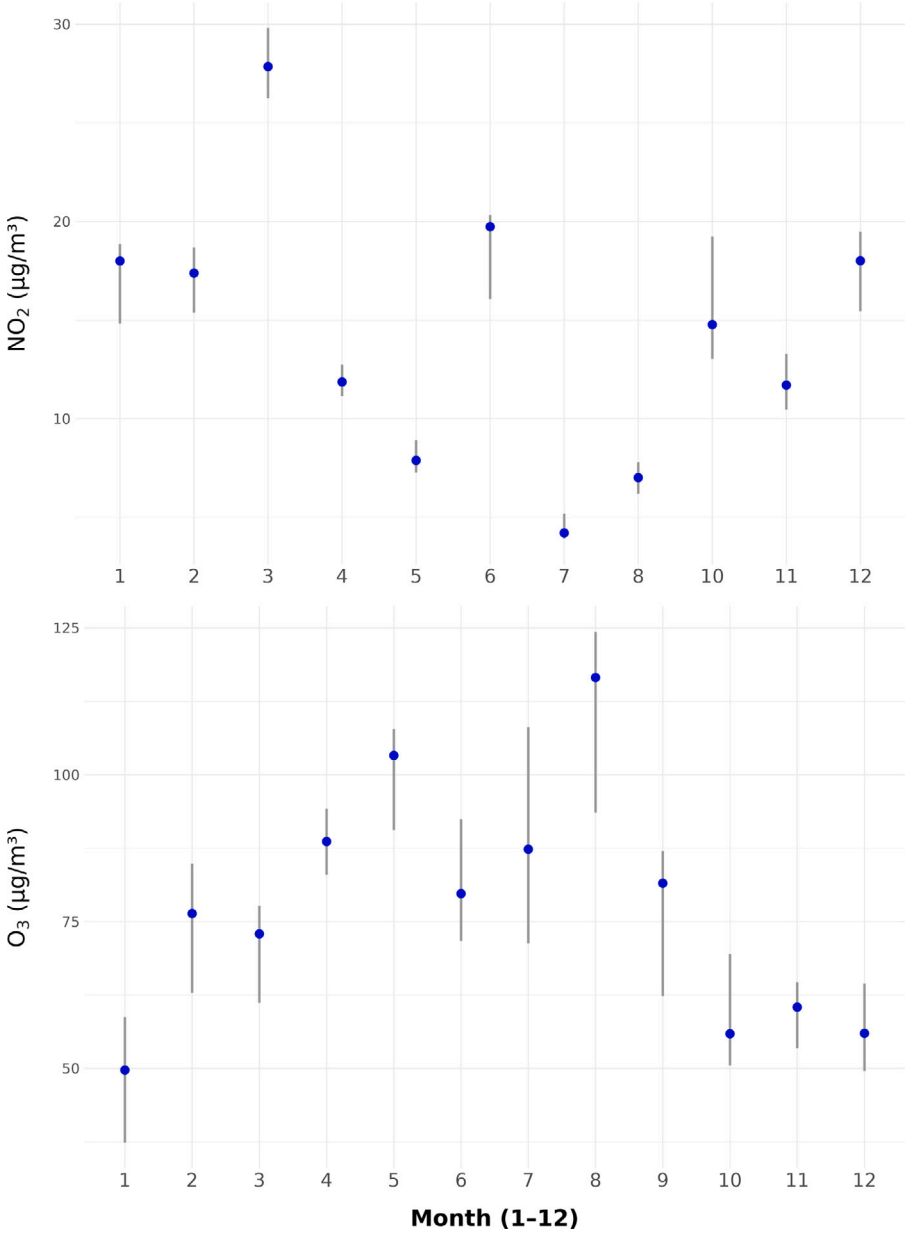

Fig. C.1. 90% Confidence intervals for NO_2 and O_3 at 12 time points.

Appendix D. Temporal variation of key predictors

To better illustrate the temporal dynamics of key features used in the model, we provide an exploratory analysis of their variation across different time scales. This is intended to complement the inclusion of temporal predictors (e.g., hour of the day, day of the week, and day of the year) discussed in Section 2.2.5.

Fig. D.1 shows the average and standard deviation of four critical variables: NO_2 concentration, number of cars (traffic intensity), temperature, and residential power consumption over hourly, weekly, and monthly timeframes.

These plots clearly reveal periodic patterns and underscore the relevance of temporal features in capturing both pollutant behavior and influencing variables. For instance, traffic and NO_2 concentrations display diurnal peaks corresponding to rush hours, weekly reductions on weekends, and seasonal trends that coincide with temperature and residential activity levels. Incorporating such features into the model enhances its ability to learn from and generalize across temporal contexts.

Data availability

The data used in this study are openly available from the official Madrid open data portal (www.datos.madrid.es), ensuring full transparency and reproducibility of the findings.

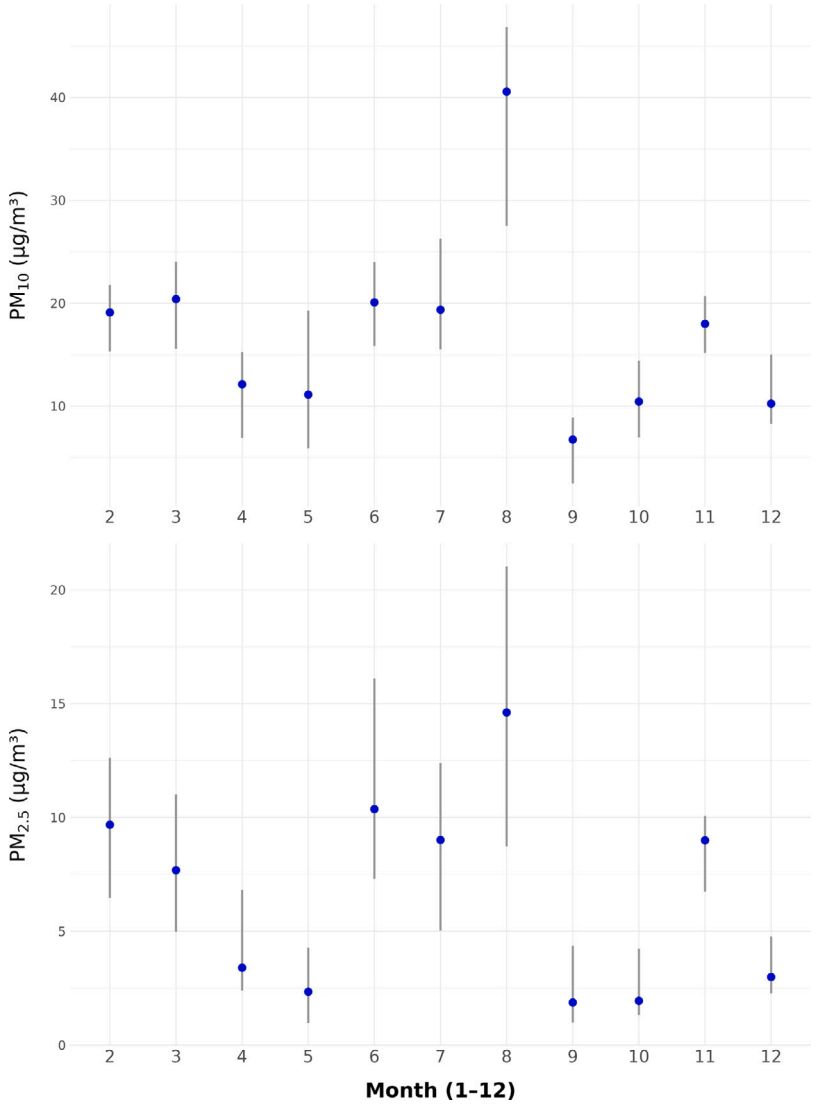

Fig. C.2. 90% Confidence intervals for $PM_{2.5}$ and PM_{10} at 12 time points.

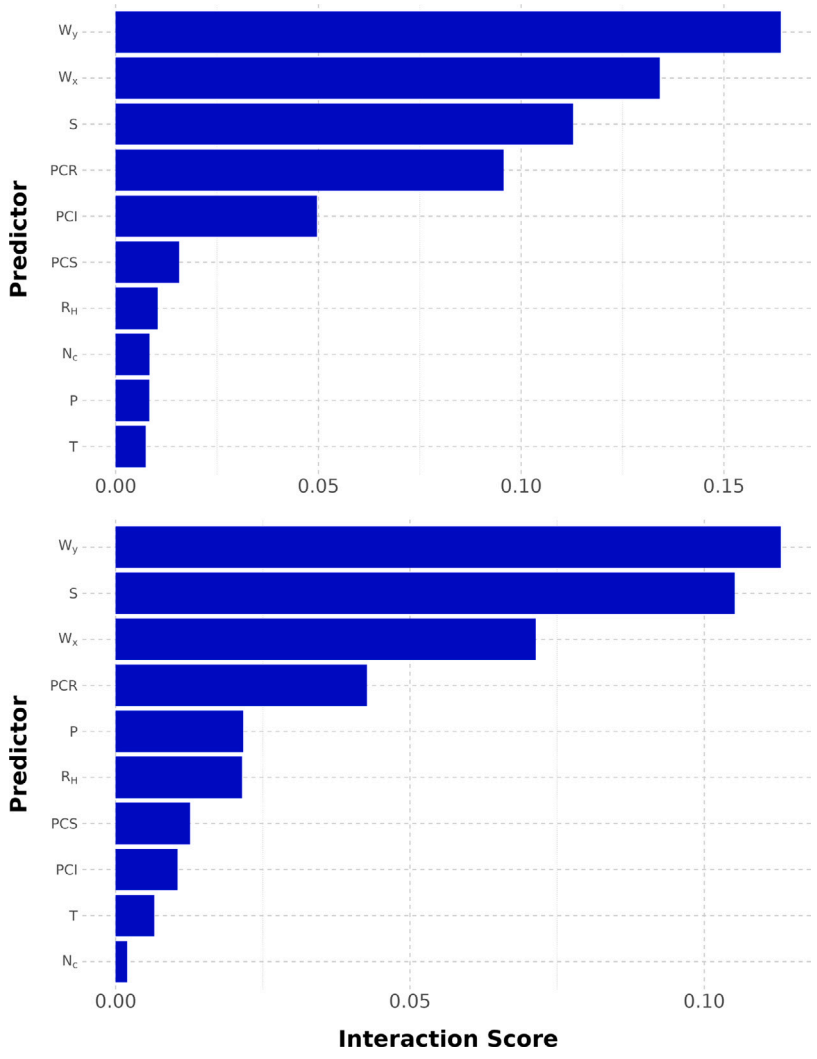

Fig. C.3. Interaction score for NO_2 (top) and O_3 (bottom).

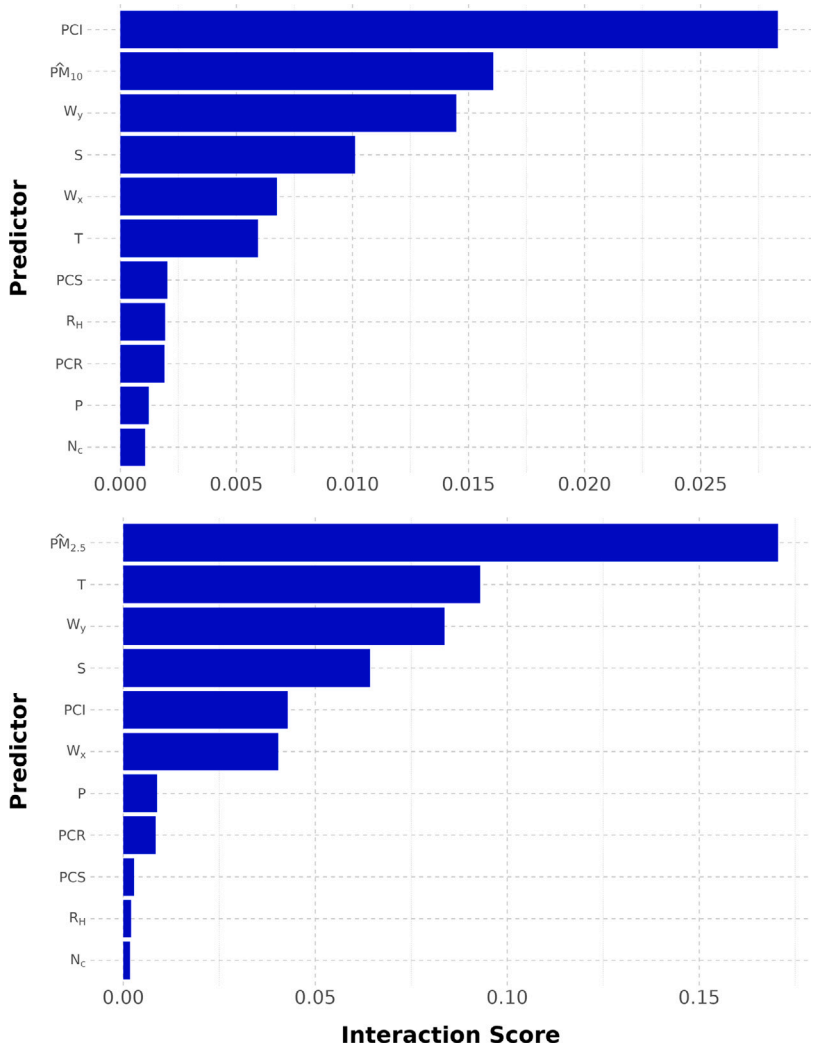

Fig. C.4. Interaction score for PM_{10} (top) and $PM_{2.5}$ (bottom).

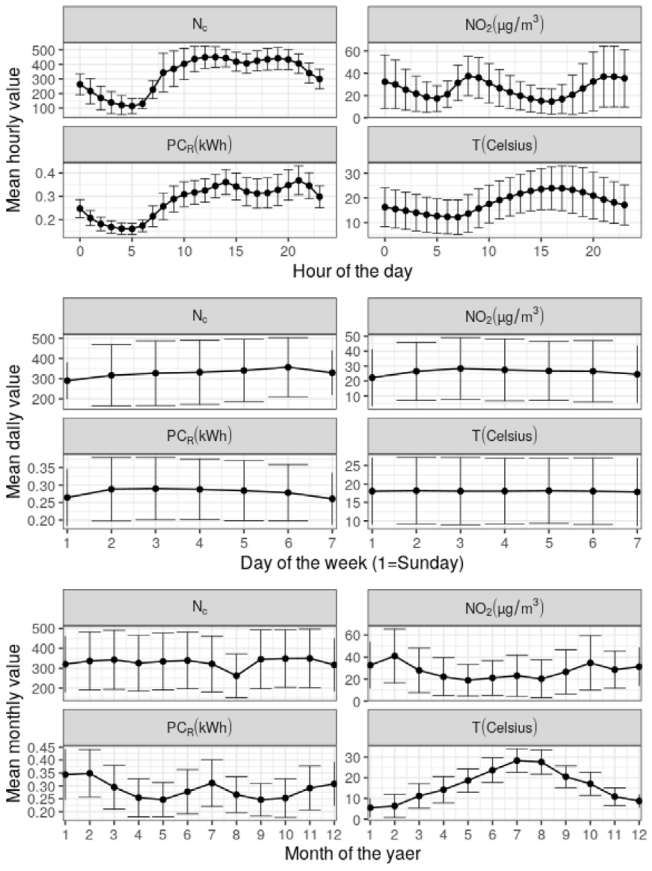

Fig. D.1. Hourly variation of NO_2, traffic volume (N_c), temperature (T), and residential power consumption (PC_R).

References

Akimoto, H., 2003. Global air quality and pollution. Science 302 (5651), 1716–1719. http://dx.doi.org/10.1126/science.1092666.

Ansari, A., Quaff, A.R., 2025. Advanced machine learning techniques for precise hourly air quality index (AQI) prediction in Azamgarh, India. Int. J. Environ. Res. 19 (1), 15.

Ayturan, Y.A., Ayturan, Z.C., Altun, H.O., 2018. Air pollution modelling with deep learning: a review. Int. J. Environ. Pollut. Environ. Model. 1 (3), 58–62.

Borge, R., Lumbreras, J., Pérez, J., de la Paz, D., Vedrenne, M., de Andrés, J.M., Rodríguez, M.E., 2014. Emission inventories and modeling requirements for the development of air quality plans. Application to Madrid (Spain). Sci. Total Environ. 466–467, 809–819. http://dx.doi.org/10.1016/j.scitotenv.2013.07.093.

Carmichael, G.R., Sandu, A., Chai, T., Daescu, D.N., Constantinescu, E.M., Tang, Y., 2008. Predicting air quality: Improvements through advanced methods to integrate models and measurements. J. Comput. Phys. 227 (7), 3540–3571. http://dx.doi.org/10.1016/j.jcp.2007.02.024, Predicting weather, climate and extreme events.

Castelli, M., Clemente, F.M., Popovič, A., Silva, S., Vanneschi, L., 2020. A machine learning approach to predict air quality in California. Complexity 2020, http://dx.doi.org/10.1155/2020/8049504.

Chelani, A.B., Gajghate, D., Tamhane, S., Hasan, M., 2001. Statistical modeling of ambient air pollutants in Delhi. Water Air Soil Pollut. 132, 315–331. http://dx.doi.org/10.1023/A:1013204120867.

de Medrano, R., de Buen Remiro, V., Aznarte, J.L., 2021. SOCAIRE: Forecasting and monitoring urban air quality in Madrid. Environ. Model. Softw. 143, 105084. http://dx.doi.org/10.1016/j.envsoft.2021.105084.

Fazakas, E., Neamtiu, I.A., Gurzau, E.S., 2024. Health effects of air pollutant mixtures (volatile organic compounds, particulate matter, sulfur and nitrogen oxides)–a review of the literature. Rev. Environ. Health 39 (3), 459–478.

García-Pérez, J., Boldo, E., Ramis, R., Pollán, M., Pérez-Gómez, B., Aragonés, N., López-Abente, G., 2007. Description of industrial pollution in Spain. BMC Public Health 7 (1), 1–13. http://dx.doi.org/10.1186/1471-2458-7-40.

Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., Boyadzhiev, D.T., 2014. Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stoch. Environ. Res. Risk Assess. 28, 1045–1060. http://dx.doi.org/10.1007/s00477-013-0800-4.

Gonçalves, M., Jiménez-Guerrero, P., Baldasano, J.M., 2009. Contribution of atmospheric processes affecting the dynamics of air pollution in South-Western Europe during a typical summertime photochemical episode. Atmospheric Chem. Phys. 9 (3), 849–864. http://dx.doi.org/10.5194/acp-9-849-2009.

Han, X., Naeher, L.P., 2006. A review of traffic-related air pollution exposure assessment studies in the developing world. Environ. Int. 32 (1), 106–120. http://dx.doi.org/10.1016/j.envint.2005.05.020.

Hassanzadeh, S., Hosseinibalam, F., Alizadeh, R., 2009. Statistical models and time series forecasting of sulfur dioxide: a case study tehran. Environ. Monit. Assess. 155, 149–155. http://dx.doi.org/10.1007/s10661-008-0424-1.

Izquierdo, R., García Dos Santos, S., Borge, R., de la Paz, D., Sarigiannis, D., Gotti, A., Boldo, E., 2020. Health impact assessment by the implementation of Madrid City air-quality plan in 2020. Environ. Res. 183, 109021. http://dx.doi.org/10.1016/j.envres.2019.109021.

Jaiswal, A., Samuel, C., Kadabgaon, V., 2018. Statistical trend analysis and forecast modeling of air pollutants. Glob. J. Environ. Sci. Manag. 4 (4), 427–438. http://dx.doi.org/10.22034/gjesm.2018.04.004.

Karnosky, D., Percy, K., Chappelka, A., Krupa, S., 2003. Air pollution and global change impacts on forest ecosystems: monitoring and research needs. Dev. Environ. Sci. 3, 447–459. http://dx.doi.org/10.1016/S1474-8177(03)03025-0.

Kim, D., Chen, Z., Zhou, L.-F., Huang, S.-X., 2018. Air pollutants and early origins of respiratory diseases. Chronic Dis. Transl. Med. 4 (2), 75–94. http://dx.doi.org/10.1016/j.cdtm.2018.03.003.

Krzyzanowski, M., Apte, J.S., Bonjour, S.P., Brauer, M., Cohen, A.J., Prüss-Ustun, A.M., 2014. Air pollution in the mega-cities. Curr. Environ. Heal. Rep. 1, 185–191. http://dx.doi.org/10.1007/s40572-014-0019-7.

Kuhn, M., Silge, J., 2022. Tidy Modeling with R. O'Reilly Media, Inc.

Kumar, U., Jain, V., 2010. ARIMA forecasting of ambient air pollutants (O 3, NO, NO 2 and CO). Stoch. Environ. Res. Risk Assess. 24, 751–760. http://dx.doi.org/10.1007/s00477-009-0361-8.

Laña, I., Del Ser, J., Padró, A., Vélez, M., Casanova-Mateo, C., 2016. The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in Madrid, Spain. Atmos. Environ. 145, 424–438. http://dx.doi.org/10.1016/j.atmosenv.2016.09.052.

Lee, B.-J., Kim, B., Lee, K., 2014. Air pollution exposure and cardiovascular disease. Toxicol. Res. 30, 71–75. http://dx.doi.org/10.5487/TR.2014.30.2.071.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525 (7569), 367–371. http://dx.doi.org/10.1038/nature15371.

Li, H., Yang, T., Du, Y., Tan, Y., Wang, Z., 2025. Interpreting hourly mass concentrations of PM2. 5 chemical components with an optimal deep-learning model. J. Environ. Sci. 151, 125–139.

Liang, Y.-C., Maimury, Y., Chen, A.H.-L., Juarez, J.R.C., 2020. Machine learning-based prediction of air quality. Appl. Sci. 10 (24), http://dx.doi.org/10.3390/app10249151.

Lin, Y.-C., Lin, Y.-T., Chen, C.-R., Lai, C.-Y., 2025. Meteorological and traffic effects on air pollutants using Bayesian networks and deep learning. J. Environ. Sci. 152, 54–70. http://dx.doi.org/10.1016/j.jes.2024.01.057.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. Front. Public Heal. 8, 14. http://dx.doi.org/10.3389/fpubh.2020.00014.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R.L., Denier van der Gon, H.A.C., Drouin, A., Elbern, H., Emili, E., Engelen, R.J., Eskes, H.J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J.W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., Ung, A., 2015. A regional air quality forecasting system over Europe: the MACC-II daily ensemble production. Geosci. Model. Dev. 8 (9), 2777–2813. http://dx.doi.org/10.5194/gmd-8-2777-2015.

Peuch, V.-H., Engelen, R., Rixen, M., Dee, D., Flemming, J., Suttie, M., Ades, M., Agustí-Panareda, A., Ananasso, C., Andersson, E., Armstrong, D., Barré, J., Bousserez, N., Dominguez, J.J., Garrigues, S., Inness, A., Jones, L., Kipling, Z., Letertre-Danczak, J., Parrington, M., Razinger, M., Ribas, R., Vermoote, S., Yang, X., Simmons, A., de Marcilla, J.G., Thépaut, J.-N., 2022. The copernicus atmosphere monitoring service: From research to operations. Bull. Am. Meteorol. Soc. 103 (12), http://dx.doi.org/10.1175/BAMS-D-21-0314.1, E2650 – E2668.

Ramanathan, V., Feng, Y., 2009. Air pollution, greenhouse gases and climate change: Global and regional perspectives. Atmos. Environ. 43 (1), 37–50. http://dx.doi.org/10.1016/j.atmosenv.2008.09.063.

Silva, R.A., West, J.J., Lamarque, J.-F., Shindell, D.T., Collins, W.J., Faluvegi, G., Folberth, G.A., Horowitz, L.W., Nagashima, T., Naik, V., et al., 2017. Future global mortality from changes in air pollution attributable to climate change. Nat. Clim. Chang. 7 (9), 647–651. http://dx.doi.org/10.1038/nclimate3354.

Wang, A., Xu, J., Tu, R., Saleh, M., Hatzopoulou, M., 2020. Potential of machine learning for prediction of traffic related air pollution. Transp. Res. Part D: Transp. Environ. 88, 102599. http://dx.doi.org/10.1016/j.trd.2020.102599.

Wang, J., Xu, W., Zhang, Y., Dong, J., 2022. A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. Chaos Solitons Fractals 158, 112098. http://dx.doi.org/10.1016/j.chaos.2022.112098.

Zhu, D., Cai, C., Yang, T., Zhou, X., 2018. A machine learning approach for air quality prediction: Model regularization and optimization. Big Data Cogn. Comput. 2 (1), http://dx.doi.org/10.3390/bdcc2010005.

Zolghadri, A., Henry, D., 2004. Minimax statistical models for air pollution time series. Application to ozone time series data measured in Bordeaux. Environ. Monit. Assess. 98, 275–294. http://dx.doi.org/10.1023/B:EMAS.0000038191.42255.7a.