environments

MDPI

Article

Towards Real-Time, High-Spatial-Resolution Air Pollution Exposure Estimation in Microenvironments Supported by Physics-Informed Machine Learning Approaches

John G. Bartzis [1,*], Ioannis A. Sakellaris [2], Spyros Andronopoulos [3], Alexandros Venetsanos [3], Fernando Martín-Llorente [4] and Stijn Janssen [5]

[1] Department of Mechanical Engineering, University of Western Macedonia, Active Urban Planning Zone (ZEP), 50100 Kozani, Greece
[2] Atmospheric Chemistry & Innovative Technologies Laboratory, Institute of Nuclear & Radiological Sciences & Technology, Energy & Safety (INRASTES), National Centre for Scientific Research "Demokritos", Aghia Paraskevi, 15310 Attiki, Greece; isakellaris@ipta.demokritos.gr
[3] Environmental Research Laboratory, Institute of Nuclear & Radiological Sciences & Technology, Energy & Safety (INRASTES), National Centre for Scientific Research "Demokritos", Aghia Paraskevi, 15310 Attiki, Greece; sandron@ipta.demokritos.gr (S.A.); venets@ipta.demokritos.gr (A.V.)
[4] CIEMAT, Research Center for Energy, Environment and Technology, Avenida Complutense 40, 28040 Madrid, Spain; fernando.martin@ciemat.es
[5] VITO NV, Flemish Institute for Research and Technology, Boeretang 200, 2400 Mol, Belgium; stijn.janssen@vito.be
* Correspondence: bartzis@uowm.gr; Tel.: +30-2461056710

Abstract

Reliable and timely estimation of air pollution exposure at high spatial and temporal resolution remains challenging in complex urban environments, where pollutant concentrations vary due to traffic emissions, urban morphology, and meteorological conditions. This study presents a physics-informed machine learning framework for near-real-time estimation of NO_2 concentrations at fine spatial scales. The approach combines a limited set of steady-state computational fluid dynamics (CFD) simulations with operational meteorological and air-quality data. CFD simulations under specific wind directions are first used to characterize site-specific dispersion patterns. These outputs are then scaled using hourly meteorological observations to generate physics-based concentration descriptors. A machine learning predictor, implemented using Random Forest and Extreme Gradient Boosting, is trained to refine these estimates by incorporating additional environmental and observational features. The method is applied to a 1 km $\times$ 1 km urban district in Antwerp, Belgium, within the FAIRMODE intercomparison framework. Validation against measurements from 105 passive samples collected over one month shows substantial improvement compared to standalone dispersion modeling, with coefficients of determination up to $R^2 = 0.965$ and reduced bias across locations. These findings demonstrate that integrating physical modeling with machine learning enables accurate and computationally efficient high-resolution exposure assessment in urban settings.

Keywords: physics-informed machine learning; urban air quality; NO_2 exposure; computational fluid dynamics (CFD); Random Forest Regression; XGBoost; microenvironment modeling; near-real-time estimation; urban dispersion; FAIRMODE

Check for updates

Academic Editors: Brindusa Sluser and Marius M. Cazacu

Received: 30 March 2026
Revised: 28 April 2026
Accepted: 30 April 2026
Published: 2 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Urban air pollution remains a major environmental and public-health challenge, particularly in densely populated urban environments where exposure to priority pollutants such as nitrogen dioxide (NO_2), particulate matter (PM), and ozone (O_3) exhibits strong spatial and temporal variability at fine scales [1–3]. Within traffic-dominated urban microenvironments, pollutant concentrations are governed by highly localized emission sources, complex street-scale flow and dispersion processes, and rapidly varying meteorological conditions, leading to pronounced concentration gradients over short distances [4,5]. Accurate characterization of population exposure in such microenvironments is therefore essential for effective air-quality management, epidemiological assessment, and the design of targeted mitigation strategies.

In response to these challenges, the European Union has established a comprehensive regulatory framework through the Ambient Air-Quality Directive (2008/50/EC), which sets legally binding limit values and target thresholds for key pollutants [6]. These objectives require Member States not only to monitor air quality but also to assess compliance, identify exceedance zones, inform the public in a timely manner, and implement effective air-quality plans when standards are not met. Achieving these objectives demands accurate, spatially resolved, and temporally responsive information on pollutant concentrations, particularly in urban microenvironments where exposure variability is highest.

Among regulated urban air pollutants, NO_2 is of particular relevance as a primary indicator of traffic-related emissions and near-road exposure [6–8]. Owing to its strong association with combustion processes and well-documented health impacts, NO_2 is widely used as a tracer for traffic-induced air pollution in urban environments [9]. Consequently, NO_2 provides a suitable first focus for the development and evaluation of high-resolution exposure estimation methodologies, while remaining representative of broader challenges associated with urban air-quality assessment.

Regulatory air-quality monitoring networks provide continuous online measurements for major priority pollutants, including NO_2, PM, and O_3, with high accuracy and data quality. However, the spatial coverage of such networks is typically limited to a relatively small number of fixed monitoring stations due to economic, infrastructural, and operational constraints [7–10]. As a result, large portions of the urban domain—particularly at the street and microenvironment scale—remain unmonitored in real time. Estimation of pollutant concentrations at unmeasured locations therefore relies on model-based spatial extrapolation approaches, including simplified dispersion models, land-use regression techniques, and urban-scale chemical transport models [7,11,12]. In complex urban environments, these methods may be associated with substantial uncertainty, especially when applied in near-real-time operational contexts [13].

High-resolution modeling frameworks, such as computational fluid dynamics (CFD) and advanced urban chemical transport models, can provide improved representation of flow, dispersion, and chemical processes, enabling the resolution of fine-scale concentration patterns for multiple pollutants [14–16]. Nevertheless, their high computational requirements generally restrict their application to offline analyses or quasi-operational scenarios, limiting their direct applicability for real-time exposure estimation across urban microenvironments.

Recent advances in Artificial Intelligence (AI) and machine learning (ML) have opened new avenues for addressing these limitations by enabling the integration of observational data with physical knowledge of atmospheric processes [17–19]. Machine learning techniques have been increasingly applied to the estimation and forecasting of urban air pollutant concentrations, including NO_2, PM, and O_3, demonstrating strong capability in capturing nonlinear relationships between emissions, meteorology, and concentrations [20–25].

However, purely data-driven approaches may lack physical consistency and robustness when applied to complex dispersion regimes. Physics-informed or physics-informed machine learning approaches, which embed physical constraints or model-derived information within data-driven frameworks, offer a promising pathway toward reliable, high-resolution, near-real-time air pollution exposure estimation in data-sparse urban environments [26,27].

The challenge here is: starting from the existing experiences of physics-informed machine learning methodologies, how we can go further to be able to address complex air-quality problems (e.g., urban environments) providing answers to the user even in real time. The recent work of Bartzis et al. [28] inaugurates such an approach looking at the specific problem of puff emission quantization in an urban microenvironment.

The present study aims to go further and develop a novel physics-informed, AI-driven methodology for near-real-time estimation of air pollution able to treat microenvironments of any complexity, by providing exposure estimations to selected locations. The proposed framework seeks applicability to the major priority pollutants (NO_2, PM, O_3), with a primary focus in this study on NO_2 as a pollutant strongly linked to traffic emissions and near-road exposure. The methodology integrates (a) historical observational data with flow and dispersion advanced modeling outputs and (b) contemporaneous online air-quality observations, typically available only at a limited number of monitoring locations. By combining physical process understanding with machine learning techniques, the proposed approach aims to provide reliable exposure estimates at any location of interest within an urban city sector. Importantly, the ambition here is to be able to support key EU air-quality objectives by enabling improved identification of pollution hotspots, enhanced exposure assessment, and the potential for real-time decision support in urban air-quality management. The methodology is demonstrated in a complex urban district in Antwerp, Belgium, within the FAIRMODE intercomparison framework, with a focus on validating its ability to deliver accurate, computationally efficient, and policy-relevant air-quality information. Ultimately, this work contributes toward bridging the gap between high-resolution scientific modeling and operational air-quality management systems required for effective implementation of EU air-quality policies. It should be noticed that this case study is drawn from the FAIRMODE intercomparison exercise [29]. This exercise was developed by the FAIRMODE WG4, which deals with the urban microscale air-quality modeling in the context of the EU Air-Quality Directives (AAQD). One of the main goals was to examine to what extent such models can be used for air-quality assessment and planning in the framework of the AAQD.

2. Methodology

2.1. The Present Site

The site selected concerns the city of Antwerp (Belgium), used in the FAIRMODE joint intercomparison exercise with the primary aim of assessing the suitability of present modeling methodologies for estimating long-term average air pollutant concentration maps in urban hotspots [29]. The selected site is an urban district of the city of Antwerp roughly 1 km $\times$ 1 km in an urban built-up area, typical of north-west European cities, consisting of a mix of street canyons and open areas. An illustration of the site is shown in Figure 1. The site data details can be found in Martín et al. (2024) [29]. A data summary is given in Table 1.

As shown in Figure 1, two monitoring stations (42R801 and 42R802) provide hourly averaged NO_2 concentrations. Additional experimental NO_2 data have been provided through an installed network of 105 passive samples within the region, collecting NO_2 during the time period from 30 April 2016 to 28 May 2016, covering almost one (1) month.

The data produced at each sample location is the NO_2 concentration averaged over the whole above period of sampling.

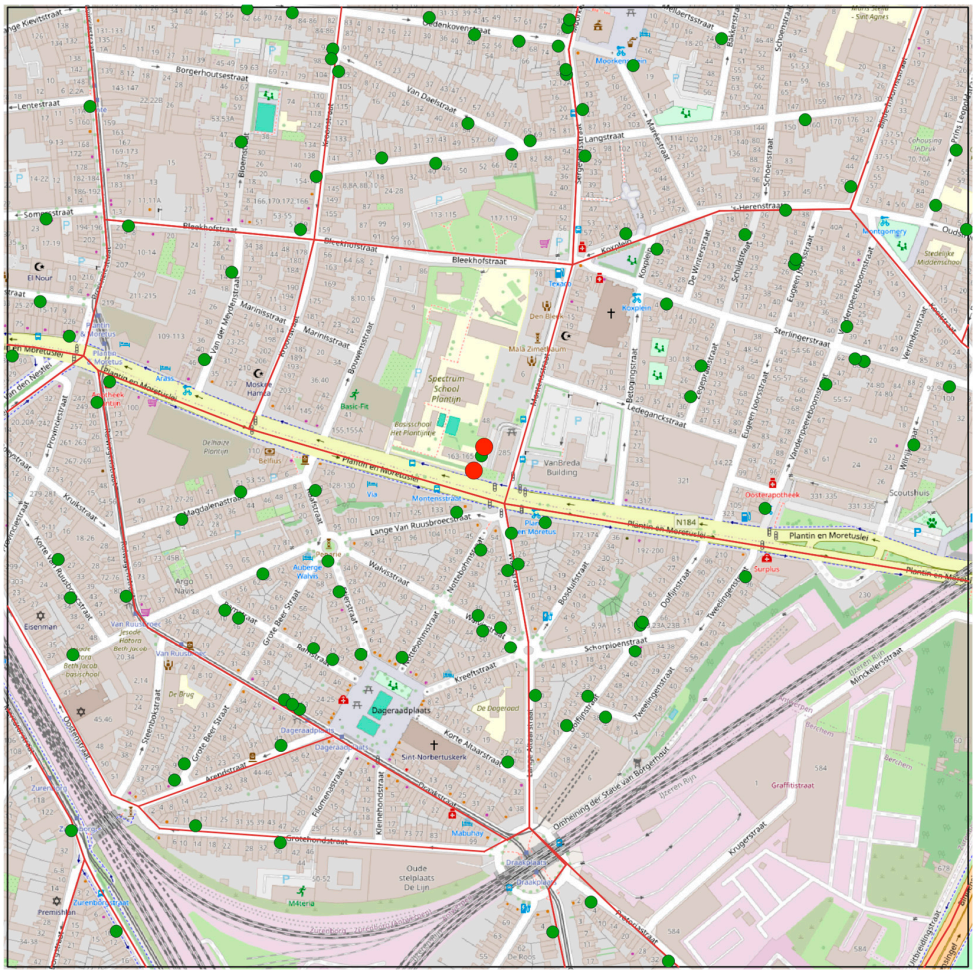


Figure 1. A Site overview. The red dots indicate the NO_2 monitoring locations and the green dots the passive sample locations.

Table 1. The site main data.

Data	Source
Measured NO_2 hourly concentrations	Two stations: Urban background (42R801); traffic station (42R802)
Modeled background NO_2 hourly concentrations (CB) representing incoming transport	The RIO model [30]
Measured concentrations with distributed passive sampling [31]	105 passive samples during period 30 April 2016 to 28 May 2016 Location range (relative to 42R801 location): Xrange: −491 m to 491 m, Yrange: −506 m to 457 m, Zrange: 2.5 m to 19 m
Measured Meteorological hourly data Wind speed and direction (30 m above ground) Near-surface (3 m) temperature, relative humidity and total radiation	One meteorological station VMM measurement station M802 at Antwerp-Luchtbal (Location: 51.261 N, 4.425 E)
Modeled annual traffic emissions for a selection of major and secondary roads	The official Flemish FASTRACE traffic emission model (version 2.1), based on COPERT 5 emission factors [32]
Modeled hourly emission factors	Provided by VITO, derived from monthly and daily profiles

2.2. The Present Approach

The basic idea is utilizing regular monitoring data to build a proper methodology able to provide near-real-time exposure estimations at any location of interest in the surrounding area. The present methodology refers to an approach where the end product is a Machine Learning Regressor (MLR) to do this task. The selected MLR is trained and tested using the available hourly data. The challenge is to establish a reliable, near-real-time operational approach. The methodology will be validated at the 105 locations where passive sampling has been carried out (Table 1).

Therefore, the proposed methodology has to remain (a) relatively simple with maximum exploitation of the existing expertise and (b) operational with minimal computer time and power. It is a stepwise approach as follows.

2.2.1. The Site Reference Simulations

A limited number of CFD simulations are carried out under reference steady flow and dispersion and a prescribed emission spatial profile. A suitable flow and dispersion model is required for this purpose. In the present study, the CFD RANS code ADREA-HF is utilized, which has been extensively used for such studies in the past [33]. For the air turbulence closure, the standard k-ε turbulence scheme has been utilized. A special characteristic of ADREA-HF is the introduction of the real urban topography into the Cartesian domain not only as volume and surface porosities but also as associated building/ground/emission surfaces through their real area and orientation. It is worth mentioning that each road has been introduced as a near-ground emitting surface.

The computation domain under study has been taken as 1020 m × 1020 m × 300 m in the west–east, south–north and vertically upwards directions respectively, including the site shown in Figure 1. The domain is divided into a 204 × 204 × 60 rectangular cell with uniform horizontal resolution 5 m × 5 m and a vertical non-uniform grid with 2 m at the ground up to 10 m at the top.

The output of CFD simulations to be utilized under the present methodology consists of steady-state NO_2 concentrations at specific locations that include (a) the two monitoring positions 42R801 and 42R802 and (b) the 105 passive sample positions mentioned above.

The limited number of CFD flow and dispersion simulations have been performed under the following conditions:

(a) All simulations are performed with a single realistic reference wind speed (VREF = 10 m/s) defined at the domain top (i.e., 300 m).

(b) Each simulation corresponds to one wind direction. Equidistant wind directions covering all 360 degrees have been selected close enough to ensure reliable spatial interpolations when needed. In the FAIRMODE intercomparison exercise, an increment of 22.5 to 45 degrees (i.e., 8–16 sectors) seems to give acceptable results. In our study, for conservative reasons, the 32 selected directions (NSIM = 32) were taken, corresponding to a wind direction increment of $\Delta\theta_{SIM} = 11.25$ degrees. Thus, the selected wind directions θ_{SIM} are as follows:

$$\theta_{SIM,ns} = (n-1)\cdot\Delta\theta_{SIM}\ (n = 1, 2 \ldots NSIM = 32)$$
$$\text{and}$$
$$\Delta\theta_{SIM} = 11.25\ degrees$$

(c) The inflow boundary condition is derived by calculating the inflow wind speed profile using ADREA-HF code, through the simulation of the corresponding 1D neutral flow with speed at the above VREF.

(d) At each road surface, the relevant traffic annual emission rate is introduced in terms of the associated emission flux ($\mu g/m^3/sec$).

The 32 CFD simulations have been performed and the computed 'reference' NO_2 concentrations CCS_{ns} (ns = 1, . . ., NSIM) at each of the 107 abovementioned locations (i.e., 105 passive sampling and two (2) pollutant station sites) have been stored to be used online, when needed.

2.2.2. The Online Flow and Dispersion Model Application

The next step is to estimate the NO_2 hourly concentrations for the year 2016 avoiding the straightforward method of direct CFD simulations. Such CFD simulations could be impractical due to demands of (a) high computational power and time and (b) additional inflow boundary conditions that could be difficult to obtain with the required accuracy. An attractive alternative would be to rely on the Step 1 reference CFD simulations and use the expertise gained up to now providing results via proper scaling/extrapolating approximations. Along this direction, a significant effort has taken place in recent years for the concentration and exposure behavior especially under high time resolution, including short-time releases in complex terrains [28]. The proposed approach has been applied to the present Antwerp data with relative success as follows:

Following Bartzis et al. [28], the hourly observed wind data given as pairs of wind speed and direction [(V_n,θ_n), n = 1, . . ., $NWIND$] have been considered. There are 8784 pairs of wind data corresponding to the number of hours of the leap year 2016. For each sensor and each hour considered, the observed wind direction θ_n is used to estimate for this specific direction, the 'reference' concentrations CSS_n derived from the abovementioned computed 'reference' CCS_{ns} via linear interpolation. Then, the present CFD-based model concentrations CSM_n are estimated, applying the 1/V scaling rule [28] as follows:

$$CSM_n = ETF_n \cdot CSS_n \cdot \frac{VREF}{V_n} \ (n = 1, \ldots, NWIND) \quad (1)$$

where ETF_n is the emission time factor given by the above Antwerp data. It is reminded that if the annual emission is multiplied by ETF_n, the resultant value is the emission considered for the specific n-hour.

It is noted that CSM_n are excess concentrations caused solely by the local emissions. To obtain the final modeled concentrations, we have to add the background concentrations (CB) given by the Antwerp data (Table 1).

$$CM_n = CB_n + CSM_n \ (n = 1, \ldots, NWIND) \quad (2)$$

where CB_n is the given hourly background concentration reflecting the contribution of the NO_2 transport outside the domain.

It is noticed that CM_n were also provided as model NO_2 concentrations to the FAIRMODE intercomparison exercise [29].

However, it is important at this stage to make the following comments:

(a) Concerning the validity of Equation (1), it is noticed that such a relationship has been successfully applied by Gaussian models [34]. In complex environments, it seems to give good results in general. In addition, the above scaling has been successfully applied to the present Antwerp data as well [35]. On the other hand, there are indications that its validity is rather confined by the validity of the CFD RANS methodology itself [36]. A deterioration is expected at specific locations where the local flow is rather unstable and the inherent sub-hourly concentration fluctuations are relatively high [36].

(b) The results depend strongly on the background concentrations, which are not always easy and straightforward to estimate.

(c) All present calculations are made assuming neutral atmospheric stability conditions. Neutral stability conditions are common in urban regions as the urban canopy complexity

often generates significant local wind velocity gradients, increasing the local turbulence production and enhancing neutral stability. However, unstable conditions are possible and cannot be ignored a priori. Preliminary data analysis has shown that in this particular case study, the ambient atmospheric stability conditions were indeed neutral for most of the hours [35].

(d) NO_2 is a rather chemically active substance and this could have an effect on the NO_2 budget, although for distances of the order of one (1) km or less, no significant effects are expected in general. On the other hand, the results of the model intercomparison exercise of FAIRMODE show that the CFD models assuming neutral stability and non-chemical reactivity yield good results compared to NO_2 observations.

(e) For completeness, one should not forget the additional errors coming from the CFD modeling assumptions including the inflow boundary conditions.

The additional consideration of AI-aided methodologies aiming to improve the results is expected to directly or indirectly remove, at least partly, some of the above drawbacks.

2.2.3. The Machine Learning Regressors (MLR)

The present focus is on suitable Supervised Machine Learning methodologies, able to handle tabulated data as already indicated above. An important property is that we are seeking methods that can meet the imposed time constraints successfully. The past experience in air pollution application is an additional criterion for selection. Among well-established regression methodologies, Random Forest Regression (RFR) and Extreme Gradient Boosting Regression (XGBR) were chosen as starting methods because they have been widely used in both meteorology and air-quality problems. RFR has been frequently applied for pollutant concentration prediction and environmental regression tasks due to its robustness and strong generalization capabilities [37,38]. Similarly, XGBR has demonstrated high predictive accuracy for air-quality estimation and related meteorological modeling [39–41]. An important characteristic of the selected models is their ability to operate efficiently under computational time constraints. Ensemble tree-based methods such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are known for their parallelizable structure and scalability. RF builds trees independently, allowing efficient parallel computation [41]. XGBoost further improves computational efficiency through optimized gradient boosting, parallel tree construction, cache-aware memory access, and sparsity-aware learning [39]. These properties make both methods particularly suitable for applications requiring rapid model training and prediction, such as environmental monitoring and operational air-quality forecasting systems.

The MLR Features

A key factor for the development of the particular MLR algorithm is the selection of the proper feature variables.

It is noticed that the site under consideration is under routine monitoring on an hourly basis that includes parameters that theoretically affect the NO_2 dispersion such as ambient wind speed and direction, temperature, humidity and solar irradiation. It has been observed that the NO_2 concentrations time series show a diurnal behavior with an integral time scale of the order of 24 h. Thus, the hour of the day is an additional feature to consider. On the other hand, in the present study, NO_2 hourly concentrations, as indicated in Table 1, are measured at two locations: 42R801 (background) and 42R802 (traffic). It is logical that the local background (42R801) concentrations should be added as an input feature.

As a consequence, the local traffic (22R802) concentrations will remain part of the target variable as it is explained below.

A critical parameter shaping flow and dispersion is ground morphology. For such a high spatial resolution, this cannot be done purely in geometrical terms. Following the conceptual frame introduced by Bartzis et al. [28] and adjusting to the present problem, the modeled concentrations' CSM, as given by Equation (1), are of high consideration. In order to avoid the involvement of the usually problematic external transport concentrations, the following input variable (XMODEL) has been proposed:

$$XMODEL = CSM - CSM_{[42R801]} + COBS_{[42R801]} \quad (3)$$

where COBS denote the measured NO_2 hourly concentrations at the two monitoring stations.

Conceptually, the XMODEL array expresses, in a quite simple way, two important determining effects: (1) the projection of the detailed ground morphology into the flow and dispersion phase through CSM and (2) the influence of the pollutant remote transport into the local domain through a single station's (42R801) observed concentrations, canceling the effect of the often problematic 'background' concentration (CB).

Concerning the target variable (TARGET) to train/test the selected MLR, the choice is rather straightforward: the observed hourly concentrations of the second station 42R802 are used, i.e.,

$$TARGET\ (training/testing) = COBS_{[42R802]} \quad (4)$$

In summary,

(a) The MLR input variables consist of the hourly arrays of (i) the observed meteostation wind speed, wind direction, temperature, relative humidity and solar irradiation, (ii) the diurnal hour (1–24 h) and (iii) the XMODEL parameter given by Equation (3).
(b) The MLR output variable designated as TARGET is the hourly array of the observed NO_2 concentrations of the second station 42R802 as indicated by Equation (4).

MLR Training and Testing

Both RFR and XGBR have been created using Python 3.9 scikit-learn. The feature matrix for training and testing for both methods is always identical. It consists of the variables described in Section "The MLR Features". Concerning the number of rows, all the hours where they have no missing data have been selected. Thus, 8342 hourly data points, out of 8784 corresponding to the year 2016, have been considered. From those data, 80% have been used for training and 20% for testing.

The results obtained for both methods give good and comparable coefficients of determination; i.e., the R^2 parameters are as follows: for the testing data R^2 = 0.936 for RFR and R^2 = 0.965 for XGBR, whereas for the training data R^2 = 0.960 for RFR and R^2 = 0.989 for XGBR.

The importance (i.e., the mean decrease in impurity and permutation importance [29] of the various input variables) is illustrated in Figure 2. The obtained values look comparable with respect to the two MLRs. The major input variable seems to be XMODEL for both RFR and XGBR. This seems quite reasonable, taking into consideration that these parameters combine morphology and key underlying physics. The wind parameters' lower importance can be explained by the fact that the wind effect, to a large degree, has already been included in XMODEL. The role of thermal-related parameters (temperature, humidity, solar radiation) seems to be limited, supporting the assumption that the neutral atmospheric stability approximation is rather valid in this specific application. RFR indicated relatively slightly higher importance for XMODEL and consequently slightly lower importance for the rest.

The present MLR behaviors with respect the testing data are illustrated in Figure 3a,b, showing the predicted vs. the observed NO_2 concentrations. The results also here look

comparable as it was expected based on the abovementioned results. The obtained results also seem quite reasonable for both methods, supporting the validity of the present concept. The XGBR predictions seem slightly better.

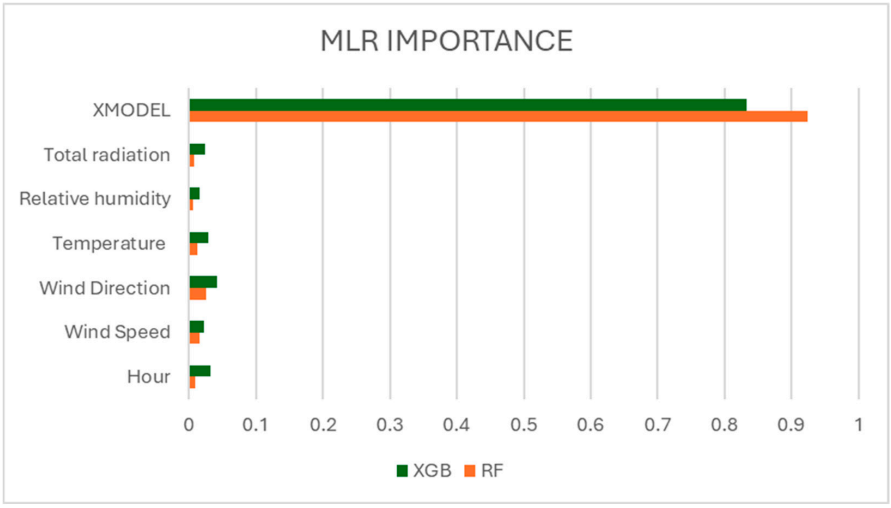


Figure 2. The present MLR input feature importance.

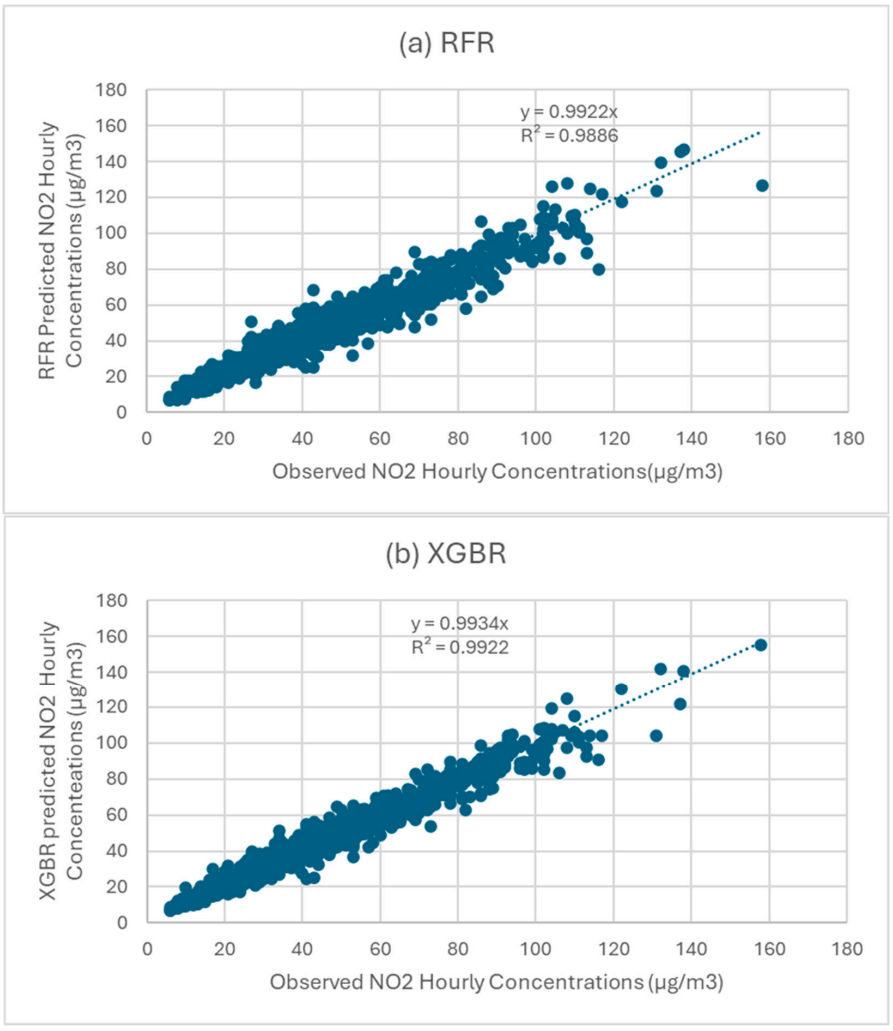


Figure 3. The present MLRs, i.e., RFR (**a**) and XGBR (**b**). Predictions for the observed NO_2 concentrations at the 42R802 station at the hours assigned for testing.

It is rather clear here that the present methodology, in reality, acts as a corrector of the already predicted pollutant concentration spatial distribution at a given hour, obtained via CFD modeling, by taking into consideration additional defining data such as (a) the meteostation observation data (i.e., temperature, relative humidity and atmospheric stability) and (b) the pollutant hourly concentration data from only one local station.

3. The Validation Exercise

The validation exercise concerns the 105 locations where the respective NO_2 mean concentrations during the period from 30 April 2016 to 28 May 2016 have been measured through passive sampling. The locations are shown in Figure 1. Concerning the geographical spread, they cover a good range both horizontally and vertically. More specifically, the horizontal range relative to 42R801 reference station is, along the X-axis, −491 m to 491 m, and the Y-axis, −506 m to 457 m. Concerning the Z-axis, the samples have been spread vertically at various heights as follows: (a) four samples near the ground (height 2.5 m to 3.5 m); (b) 85 samples at the first floor (height 3.23–7.6 m); (c) 11 samples at the second floor (height 4.34 to 10.85 m); (d) five at the third floor (height 10.6 m); (e) one (1) sample at the sixth floor (height 19 m).

Each MLR is used to provide NO_2 hourly concentrations at the 105 locations during the above sampling period (i.e., 696 h) and then, the averages for the campaign period were computed. These averages are compared with the observed ones seeking for agreement.

Looking at the data during this period, it is noticed that several NO_2 hourly concentration data points from the selected reference station 42R801 were missing. In fact, 31 values out of 696 were missing. A respective imputation process was needed to be applied. Keeping things simple and practical, an imputation process has been performed using the RFR approach as follows: The NO_2 hourly concentrations at the 42R801 station now become the target variable. Concerning the input variables, the abovementioned meteostation observation data, as well as the diurnal hour, are also included here. The additional input parameter is the given NO_2 background hourly concentrations (Table 1), a measure of NO_2 concentration time profiling. A new RFR predictor has been built using that data. The same training/testing strategy (i.e., 80/20) has been followed. The results obtained gave good R^2 statistics: R^2 = 0.912 for the testing data and R^2 = 0.945 for the training data. The predicted missing concentrations have been added to the 42R801 NO_2 concentrations in order to be able to build a complete array to be used for the present validation exercise. It is noticed that only one method (i.e., the RF method) has been applied since the intention was to use an identical dataset for both methods used for the validation phase.

The core validation exercise concerns the performance of the selected two MLRs (i.e., RFR and XGBR) in predicting the average NO_2 concentrations at the 105 sampling sites; the results for each sample are illustrated in Figure 4. The samples are presented in concentration size order. It is interesting to note that both MLRs gave almost identical results. It is clear that significant improvements have been obtained by MLRs. More specifically, the agreement between MLR predictions with observations at the locations with moderate concentrations is remarkable. Concerning the elevated concentrations, the predictions follow the observation trends. The comparisons here also seem to be reasonable, if one takes into consideration that the locations with relatively high concentrations are often exposed to irregular emissions of highly local character. Such improvements are also illustrated on global terms, as in Figure 5 showing the average concentration comparisons over all 105 samples. On the other hand, the MLRs have decisively corrected the CFD model predictions. The CFD underprediction needs to be further looked at as to what degree it is affected by the specific CB array input. Additionally, in the high-exposure regime, the predicted peaks follow the real trend. Concerning the discrepancies, the predictions seem

to follow the flow and dispersion model trend as expected, since the MLRs are trained by the model.

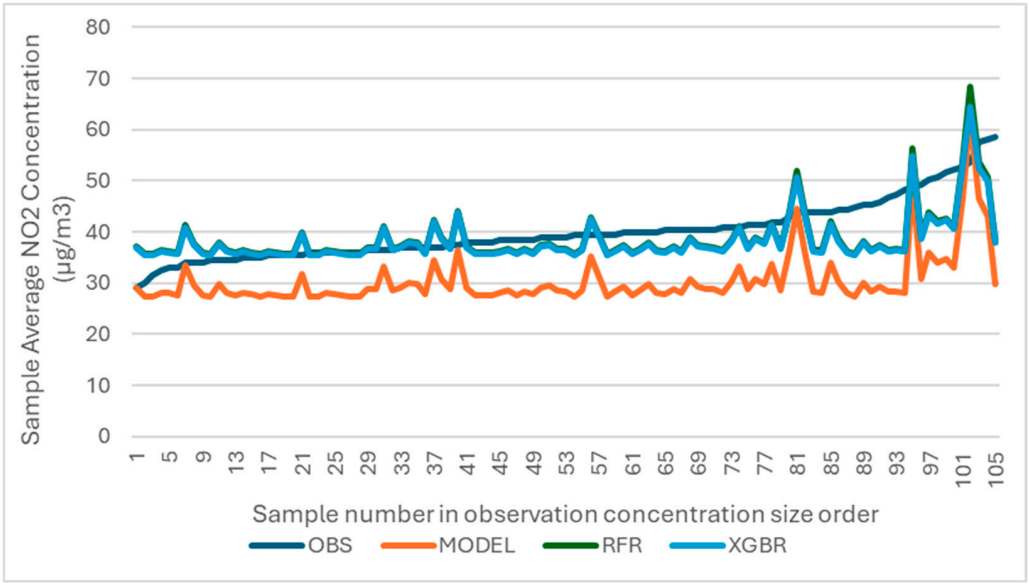


Figure 4. The present MLR (i.e., RFR and XGBR) predictions for the observed average NO_2 concentrations measured by each passive sample. The sample numbering has been made according to the measured average concentration value starting from the lowest one.

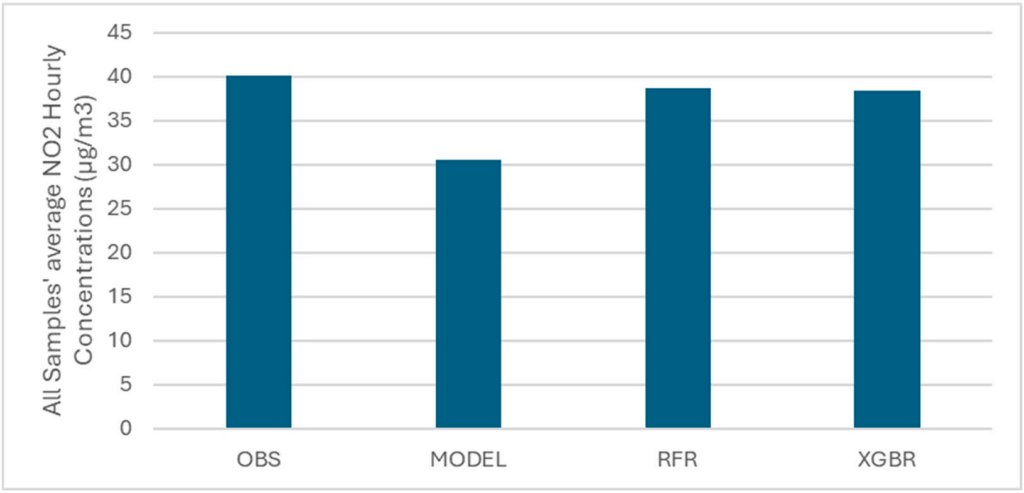


Figure 5. The present MLR (i.e., RFR and XGBR) predictions for the observed NO_2 concentrations averaged over all samples.

It is interesting to see to what degree the height of the sampling location affects the performance of the predictions. Table 2 gives an indication showing the average concentrations per floor. It is clear that the MLRs improve considerably. The most significant discrepancy appears at the sixth floor measurement and could be caused by non-local transport, which becomes more important at higher altitudes; however, the discrepancy should not be given too much weight because there is only one sample at that elevation.

Table 2. Average sample NO_2 concentrations at various vertical positions ($\mu g/m^3$).

Floor	No of Samples	OBS	MODEL	RFR	XGBR
Near Ground	4	41.61	29.16	37.35	37.09
1st	84	40.18	30.53	38.64	38.27
2nd	11	38.38	30.46	38.57	38.28
3rd	5	40.58	32.26	40.33	39.91
6th	1	44.78	27.37	35.70	35.42

4. Concluding Remarks

This study introduced a physics-informed, AI-driven methodology for near-real-time estimation of NO_2 concentrations in complex urban microenvironments. The proposed approach combines a limited but physically consistent set of CFD reference simulations with operational meteorological and monitoring data, and with MLRS as corrective and adaptive components. The key innovation lies in embedding morphology- and physics-based descriptors (XMODEL) into the feature space of the machine learning predictor, thereby preserving the fundamental dispersion behavior while allowing data-driven correction of residual errors.

Application to the Antwerp FAIRMODE case study demonstrated that both Random Forest Regression and Extreme Gradient Boosting significantly improve concentration estimates compared to pure dispersion modeling. The improvement is consistent across most sampling heights and spatial locations, with the largest discrepancies appearing at higher elevations where non-local transport effects become more dominant. The results confirm that the proposed framework effectively acts as a physically guided model corrector, capable of capturing additional influences such as atmospheric stability and secondary effects without requiring computationally intensive real-time CFD simulations.

The methodology satisfies the two principal operational requirements: (a) computational efficiency suitable for near-real-time applications and (b) maximum exploitation of existing physical modeling expertise. Although the present study focused on NO_2 as a traffic-related tracer pollutant, its conceptual approach can be used as a basis for confronting the problem related to other priority pollutants such as PM and O_3, provided that appropriate background treatment and chemical considerations are incorporated.

It should also be reiterated that the CFD-based modeling has been considerably simplified. Such simplifications include (a) neutral atmospheric conditions as well as omission of the NO_2 chemistry and (b) application of the RANS approach and wind speed scaling (Equation (1)) even in locations that might have an unstable local flow behavior or/and significant fluctuations of the sub-hourly concentrations. Further research is needed on how to cope successfully with those problems, touching probably the CFD RANS methodology itself, in practical terms. It should be underlined here that the present AI-based methodology is expected to tackle, in a corrective way, the above weaknesses.

Future work should also focus on systematic multi-site validation, incorporation of stability-dependent corrections and refinement of background concentration treatment.

Integration within operational air-quality management systems and smart-city platforms represents a natural next step. It is underlined that the application of the present methodology could provide near-real-time answers to the manager/operator in charge without the need for excessive computational facilities allowing for timely interventions. Overall, the study provides a solid methodological foundation toward reliable, high-resolution, real-time urban exposure estimation through physics-informed machine learning.

Author Contributions: Conceptualization, J.G.B.; Methodology, J.G.B., I.A.S., S.A. and A.V.; Software, J.G.B., I.A.S., S.A. and A.V.; Validation, J.G.B., S.A. and F.M.-L.; Formal analysis, J.G.B., I.A.S. and S.A.; Investigation, J.G.B.; Data curation, J.G.B., F.M.-L. and S.J.; Writing—original draft preparation, J.G.B. and I.A.S. Writing—review and editing, All; Visualization, I.A.S.; Supervision, J.G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors are thankful for participation in the FAIRMODE Microscale Modeling Intercomparison Exercise.

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

Abbreviations

The following abbreviations are used in this manuscript:

NO_2	Nitrogen Dioxide
PM	Particulate Matter
O_3	Ozone
CFD	Computational Fluid Dynamics
AI	Artificial Intelligence
ML	Machine Learning
FAIRMODE	Forum for Air Quality Modeling
MLR	Machine Learning Regressors
RFR	Random Forest Regression
XGBR	Extreme Gradient Boosting Regression
RF	Random Forest
XGBoost	Extreme Gradient Boosting

References

1. Beelen, R.; Raaschou-Nielsen, O.; Stafoggia, M.; Andersen, Z.J.; Weinmayr, G.; Hoffmann, B.; Wolf, K.; Samoli, E.; Fischer, P.; Nieuwenhuijsen, M.; et al. Effects of Long-Term Exposure to Air Pollution on Natural-Cause Mortality: An Analysis of 22 European Cohorts within the Multicentre ESCAPE Project. *Lancet* **2014**, *383*, 785–795. [CrossRef]
2. Targa, J.; Colina, M.; Banyuls, L.; González Ortiz, A.; Soares, J. *Status Report of Air Quality in Europe for Year 2023, Using Validated Data (ETC HE Report 2025/2)*; Zenodo: Geneva, Switzerland, 2025.
3. WHO. *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; WHO Guidelines Approved by the Guidelines Review Committee; World Health Organization: Geneva, Switzerland, 2021; ISBN 978-92-4-003422-8.
4. Soulhac, L.; Salizzoni, P.; Cierco, F.-X.; Perkins, R. The Model SIRANE for Atmospheric Urban Pollutant Dispersion; Part I, Presentation of the Model. *Atmos. Environ.* **2011**, *45*, 7379–7395. [CrossRef]
5. Vardoulakis, S.; Fisher, B.E.A.; Pericleous, K.; Gonzalez-Flesca, N. Modelling Air Quality in Street Canyons: A Review. *Atmos. Environ.* **2003**, *37*, 155–182. [CrossRef]
6. European Parliament. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. *Off. J. Eur. Union* **2008**, *152*, 1–44.
7. Berkowicz, R. OSPM—A Parameterised Street Pollution Model. *Environ. Monit. Assess.* **2000**, *65*, 323–331. [CrossRef]
8. Snyder, E.G.; Watkins, T.H.; Solomon, P.A.; Thoma, E.D.; Williams, R.W.; Hagler, G.S.W.; Shelow, D.; Hindin, D.A.; Kilaru, V.J.; Preuss, P.W. The Changing Paradigm of Air Pollution Monitoring. *Environ. Sci. Technol.* **2013**, *47*, 11369–11377. [CrossRef]
9. Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A Review of Land-Use Regression Models to Assess Spatial Variation of Outdoor Air Pollution. *Atmos. Environ.* **2008**, *42*, 7561–7578. [CrossRef]
10. Zhang, Y.; Bocquet, M.; Mallet, V.; Seigneur, C.; Baklanov, A. Real-Time Air Quality Forecasting, Part I: History, Techniques, and Current Status. *Atmos. Environ.* **2012**, *60*, 632–655. [CrossRef]
11. Lateb, M.; Meroney, R.N.; Yataghene, M.; Fellouah, H.; Saleh, F.; Boufadel, M.C. On the Use of Numerical Modelling for Near-Field Pollutant Dispersion in Urban Environments—A Review. *Environ. Pollut.* **2016**, *208*, 271–283. [CrossRef]

12. Tominaga, Y.; Stathopoulos, T. CFD Simulation of Near-Field Pollutant Dispersion in the Urban Environment: A Review of Current Modeling Techniques. *Atmos. Environ.* **2013**, *79*, 716–730. [CrossRef]
13. Baklanov, A.; Schlünzen, K.; Suppan, P.; Baldasano, J.; Brunner, D.; Aksoyoglu, S.; Carmichael, G.; Douros, J.; Flemming, J.; Forkel, R.; et al. Online Coupled Regional Meteorology Chemistry Models in Europe: Current Status and Prospects. *Atmos. Chem. Phys.* **2014**, *14*, 317–398. [CrossRef]
14. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
15. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Prabhat Deep Learning and Process Understanding for Data-Driven Earth System Science. *Nature* **2019**, *566*, 195–204. [CrossRef]
16. Wang, R.; Yu, R. Physics-Guided Deep Learning for Dynamical Systems: A Survey. *arXiv* **2023**, arXiv:2107.01272. [CrossRef]
17. de Villeroché, A.; Le Guen, V.; Mouradi, R.-S.; Massin, P.; Bocquet, M.; Farchi, A.; Cheng, S.; Armand, P. Physics-Informed Neural Networks for Atmospheric Flow Modeling of Pollutant Dispersion in Industrial Sites. *Air Qual. Atmos. Health* **2026**, *19*, 38. [CrossRef]
18. Grange, S.K.; Carslaw, D.C.; Lewis, A.C.; Boleti, E.; Hueglin, C. Random Forest Meteorological Normalisation Models for Swiss PM_{10} Trend Analysis. *Atmos. Chem. Phys.* **2018**, *18*, 6223–6239. [CrossRef]
19. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *arXiv* **2022**, arXiv:2003.04919. [CrossRef]
20. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An Ensemble-Based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution. *Environ. Int.* **2019**, *130*, 104909. [CrossRef]
21. Li, P.; Zhang, T.; Jin, Y. A Spatio-Temporal Graph Convolutional Network for Air Quality Prediction. *Sustainability* **2023**, *15*, 7624. [CrossRef]
22. Liao, Q.; Zhu, M.; Wu, L.; Pan, X.; Tang, X.; Wang, Z. Deep Learning for Air Quality Forecasts: A Review. *Curr. Pollut. Rep.* **2020**, *6*, 399–409. [CrossRef]
23. Kheder, A.; Toropainen, H.; Peng, W.; Antão, S.; Chen, J.; Boy, M.; Liu, Z.-S. TopoFlow: Topography-Aware Pollutant Flow Learning for High-Resolution Air Quality Prediction. *npj Clim. Atmos. Sci.* **2026**. [CrossRef]
24. Jurado, X.; Reiminger, N.; Maurer, L.; Vazquez, J.; Wemmert, C. Assessment of a Deep Learning Model for Monitoring Atmospheric Pollution: Case Study in Antwerp, Belgium. *Sustain. Cities Soc.* **2023**, *99*, 104951. [CrossRef]
25. Jurado, X.; Reiminger, N.; Benmoussa, M.; Vazquez, J.; Wemmert, C. Deep Learning Methods Evaluation to Predict Air Quality Based on Computational Fluid Dynamics. *Expert Syst. Appl.* **2022**, *203*, 117294. [CrossRef]
26. Anitescu, C.; İsmail Ateş, B.; Rabczuk, T. Physics-Informed Neural Networks: Theory and Applications. In *Machine Learning in Modeling and Simulation: Methods and Applications*; Rabczuk, T., Bathe, K.-J., Eds.; Springer International Publishing: Cham, Switzerland, 2023; pp. 179–218. ISBN 978-3-031-36644-4.
27. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-Informed Machine Learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
28. Bartzis, J.; Andronopoulos, S.; Sakellaris, I. Source Term Estimation for Puff Releases Using Machine Learning: A Case Study. *Atmosphere* **2025**, *16*, 697. [CrossRef]
29. Martín, F.; Janssen, S.; Rodrigues, V.; Sousa, J.; Santiago, J.L.; Rivas, E.; Stocker, J.; Jackson, R.; Russo, F.; Villani, M.G.; et al. Using Dispersion Models at Microscale to Assess Long-Term Air Pollution in Urban Hot Spots: A FAIRMODE Joint Intercomparison Exercise for a Case Study in Antwerp. *Sci. Total Environ.* **2024**, *925*, 171761. [CrossRef]
30. Janssen, S.; Dumont, G.; Fierens, F.; Mensink, C. Spatial Interpolation of Air Pollution Measurements Using CORINE Land Cover Data. *Atmos. Environ.* **2008**, *42*, 4884–4903. [CrossRef]
31. De Craemer, S.; Vercauteren, J.; Fierens, F.; Lefebvre, W.; Meysman, F.J.R. Using Large-Scale NO_2 Data from Citizen Science for Air-Quality Compliance and Policy Support. *Environ. Sci. Technol.* **2020**, *54*, 11070–11078. [CrossRef]
32. Degraeuwe, B.; Hooyberghs, H.; Janssen, S.; Lefebvre, W.; Maiheu, B.; Megaritis, A.; Vanhulsel, M. A Source Apportionment and Air Quality Planning Methodology for NO_2 Pollution from Traffic and Other Sources. *Environ. Model. Softw.* **2024**, *176*, 106032. [CrossRef]
33. Andronopoulos, S.; Bartzis, J.G.; Würtz, J.; Asimakopoulos, D. Modelling the Effects of Obstacles on the Dispersion of Denser-than-Air Gases. *J. Hazard. Mater.* **1994**, *37*, 327–352. [CrossRef]
34. Hanna, S.R.; Briggs, G.A.; Hosker, J. *Handbook on Atmospheric Diffusion*; National Oceanic and Atmospheric Administration: Oak Ridge, TN, USA; Atmospheric Turbulence and Diffusion Lab.: Oak Ridge, TN, USA, 1982.
35. Bartzis, J.; Sakellaris, I.; Tolias, I.; Venetsanos, A. Simplified microscale modeling methodologies for urban air quality. In *Proceedings of Abstracts 13th International Conference on Air Quality: Science and Application*; University of Hertfordshire: Hatfield, UK, 2022.

36. Bartzis, J.G.; Efthimiou, G.C.; Andronopoulos, S. Modelling Exposure from Airborne Hazardous Short-Duration Releases in Urban Environments. *Atmosphere* **2021**, *12*, 130. [CrossRef]
37. Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
38. Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M.L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal Prediction of Continuous Daily PM2.5 Concentrations across China Using a Spatially Explicit Machine Learning Algorithm. *Atmos. Environ.* **2017**, *155*, 129–139. [CrossRef]
39. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
40. Zhan, Y.; Luo, Y.; Deng, X.; Grieneisen, M.L.; Zhang, M.; Di, B. Spatiotemporal Prediction of Daily Ambient Ozone Levels across China Using Random Forest for Human Exposure Assessment. *Environ. Pollut.* **2018**, *233*, 464–473. [CrossRef]
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.