Environmental Science & Technology

pubs.acs.org/est

Article

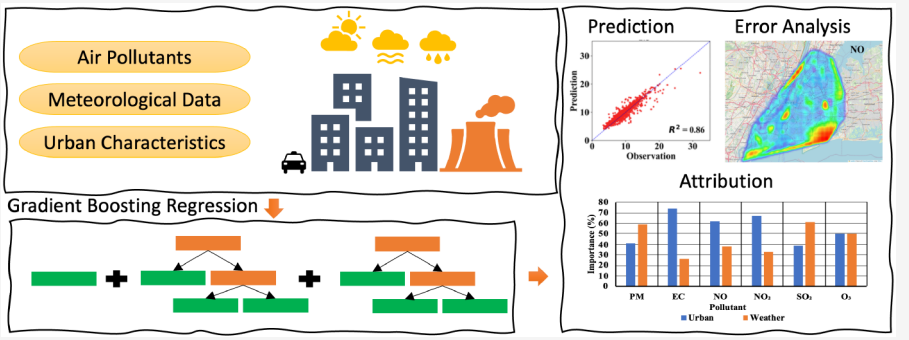
Weather and the City: Machine Learning for Predicting and Attributing Fine Scale Air Quality to Meteorological and Urban Determinants

Firas Gerges, Maider Llaguno-Munitxa, Mark A. Zondlo, Michel C. Boufadel, and Elie Bou-Zeid*

Cite This: *Environ. Sci. Technol.* 2024, 58, 6313−6325

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Urban air quality persists as a global concern, with critical health implications. This study employs a combination of machine learning (gradient boosting regression, GBR) and spatial analysis to better understand the key drivers behind air pollution and its prediction and mitigation strategies. Focusing on New York City as a representative urban area, we investigate the interplay between urban characteristics and weather factors, showing that urban features, including traffic-related parameters and urban morphology, emerge as crucial predictors for pollutants closely associated with vehicular emissions, such as elemental carbon (EC) and nitrogen oxides (NO_x). Conversely, pollutants with secondary formation pathways (e.g., $PM_{2.5}$) or stemming from nontraffic sources (e.g., sulfur dioxide, SO_2) are predominantly influenced by meteorological conditions, particularly wind speed and maximum daily temperature. Urban characteristics are shown to act over spatial scales of 500 × 500 m^2, which is thus the footprint needed to effectively capture the impact of urban form, fabric, and function. Our spatial predictive model, needing only meteorological and urban inputs, achieves promising results with mean absolute errors ranging from 8 to 32% when using full-year data. Our approach also yields good performance when applied to the temporal mapping of spatial pollutant variability. Our findings highlight the interacting roles of urban characteristics and weather conditions and can inform urban planning, design, and policy.

KEYWORDS: *air quality, machine learning, urban environment, spatial analysis, pollution modeling*

1. INTRODUCTION

Air pollution remains a complex challenge with far-reaching implications for public health and well-being, not only in low-income countries, but also in middle- and high-income ones.[1−6] In many cities, air pollution levels exceed World Health Organization (WHO) guidelines[7] calling for new and innovative approaches to better understand the sources and drivers of air pollutants. Although this hazard affects both urban and rural residents, the former are particularly vulnerable due to higher population densities,[8,9] higher traffic volumes,[10−12] slower ventilation due to urban form,[13,14] and increased industrial and commercial activities that contribute to higher emissions.[15]

While the mechanistic links are qualitatively evident, the exact relation between air pollutant concentrations and these urban characteristics is not yet thoroughly understood and quantified, hindering effective policies and initiatives to combat air pollution in urban environments. One uncontested challenge is the complexity and nonlinearity of the linkages, but a lesser appreciated difficulty is grasping the spatial scale over which urban emissions and characteristics should be considered. Understanding this "footprint scale" of urban features, as they relate to different air pollutants, is essential since different urban characteristics have varied spatial scales of

Received: January 22, 2024
Revised: March 13, 2024
Accepted: March 14, 2024
Published: March 26, 2024

influence, and their effect on pollutant concentrations at a point depends on an "environmental neighborhood" around that location.[16] A closely related challenge is comparing the influence of the city and its properties vis-à-vis meteorological drivers in modulating spatial and temporal pollution variability. Therefore, innovative approaches to determining the spatial connections between urban characteristics and air pollutant concentrations, coupled with information about weather patterns, can help identify the primary drivers of elevated pollutant concentrations. This knowledge can then be applied to develop more effective and rapid fine-scale prediction, mitigation, and management frameworks for the improvement of urban air quality. Such data-driven models can provide high-resolution short-term forecasts that complement more classical, physics-based approaches. Physical models such as the Weather Research and Forecasting model coupled with Chemistry (WRF-CHEM) or the Community Multiscale Air Quality Modeling (CMAQ) are typically applied with resolutions >1 km. They can thus predict the long-term coarse trends of urban air quality but cannot provide the fine scale maps needed for various decisions by the residents, planners, and managers of cities.

Several previous studies have already investigated the impact of urban parameters on air quality on fine scales. Li et al.[17] examined the spatial variations of several air pollutants using quantifiable metrics representing the urban environment, demonstrating that distance to roads and buildings are among the features with most influence over the spatial variations of air pollution. Yuan et al.[18] investigated the impact of urban permeability and building geometries on air pollutant dispersion using numerical simulations. The authors highlighted the importance of understanding the relationship between urban morphology and air pollutant dispersion in high-density cities and suggested that changes in urban design and planning can be effective strategies for improving air quality in these areas. Llaguno-Munitxa and Bou-Zeid[16] introduced the concept of an environmental neighborhood to analyze the spatial extent of the effects of several urban parameters on air quality. The authors applied linear correlation analysis to urban features−pollutant pairs to analyze their relations in New York City and found that the scale of environmental neighborhoods (determined as the scale that maximizes the correlation) ranges from approximately 200 to 1000 m, depending on the urban parameter and pollutant being considered. What remains elusive, however, is an understanding of the relative importance of these urban drivers versus meteorological determinants of air quality and how both can be integrated to better predict air pollution in cities.

Machine learning (ML) provides powerful tools to address this gap and to predict complex data sets across a wide range of fields; it is likely to be more capable of elucidating urban air pollution drivers than linear correlations. One of the key advantages of ML is its ability to identify patterns and relationships, linear or otherwise, within data that may be difficult or nearly impossible to detect using traditional statistical methods.[19] This is particularly relevant in the context of air quality research, where the relationships between various meteorological and urban features and pollutant concentrations are complex and not easily quantifiable. ML algorithms can be used to map and analyze air pollution,[20−23] as well as to identify key predictors of air pollutant concentrations from data sets of meteorological data, traffic patterns, and urban features. Some of the recent applications of ML to urban air quality include the work of Song et al.[24] who used machine learning, in particular, Gradient Boosting, to produce city-scale $PM_{2.5}$ maps from mobile sensing and urban big data. Chai et al.[25] employed XGBoost to predict the spatial distribution of $PM_{2.5}$ using meteorological and land use data. Li et al.[26] used Random Forest to predict the concentration of $PM_{2.5}$ and coupled it with permutation importance and partial dependence plots to quantify the contribution of various chemical and elemental composition to air pollution events. Findings showed that $PM_{2.5}$ is mostly sensitive to NH_4^+, NO_3^-, and SO_4^{2-}, in that order. Zhang et al.[27] quantified anthropogenic and weather drivers of $PM_{2.5}$ concentrations using Random Forest; the authors attributed an improved air quality in terms of $PM_{2.5}$ to a decrease in anthropogenic emission. The importance of weather drivers varied between locations, where zonal wind speed at 500 hpa, relative humidity, and total precipitation were among the dominant factors. It is important to note that alongside machine learning, other empirical, statistical models such as land-use regression models have also been widely employed in air quality research.[28−30]

In this study, we also aim to leverage machine learning with spatial analysis to predict and analyze critical air pollutants, but our approach is distinct from previous literature by (i) its embedding of environmental footprint analysis within the ML framework, (ii) the concurrent use of ML for error analysis (not only for prediction and attribution), and (iii) a focus on attribution analysis aimed at examining what pollutants are more affected by urban drivers and what pollutants are more strongly modulated by the weather. In particular, we leverage gradient boosting regression (GBR) to generate preliminary feature importance and determine the most critical footprint scale for each feature group. Furthermore, we perform feature importance and attribution analyses, which allow us to identify the most important factors that contribute to air pollution in our urban area. Subsequently, we expand the analysis to model the errors in our initial predictions, gaining insight into the factors that may result in poor model performance and large differences in prediction error of pollutant concentrations across different locations. This could be particularly valuable in identifying and addressing sources of error that may be specific to certain areas or conditions, which can prove valuable in informing policies and interventions aimed at mitigating the impact of air pollution on public health and the environment. Furthermore, we perform additional experiments to forecast the spatial maps of air quality through time. Our pollutants include particulate matter ($PM_{2.5}$), nitrogen oxides (NO_x), sulfur dioxide (SO_2), ozone (O_3), and elemental carbon (EC), in the highly dense urban setting of New York City.

We address the following research questions:

- At what spatial scales should urban features be integrated to predict air quality, and how do these scales differ between pollutants?
- How do weather patterns, traffic, and urban characteristics modulate air pollution concentrations, and are there distinct drivers for distinct pollutants?
- How do prediction errors for air pollutant concentrations differ between monitoring sites, and what are the urban and climate drivers of larger errors?
- How effective is machine learning, and gradient boosting regression, in particular, in predicting and mapping the

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

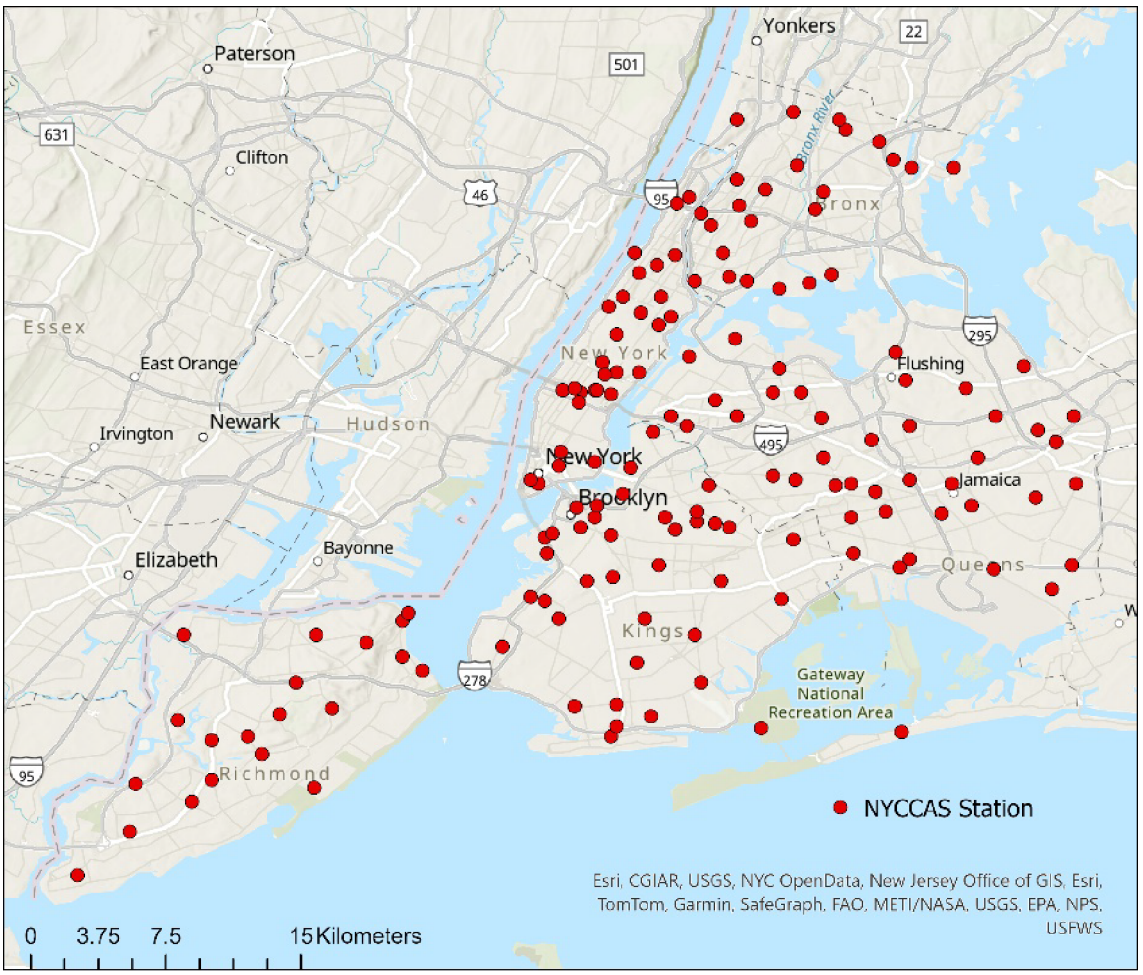


Figure 1. Locations of the used NYCCAS air quality stations (red circles) within NYC. Map created by the authors using the ArcGIS Pro software.

spatial variability of critical air pollutants in urban settings?

2. MATERIALS AND METHODS

The primary objectives of the study are to leverage machine learning for predicting and analyzing critical air pollutants in urban settings, understanding the impact of urban features on air quality, and investigating the interplay among meteorological conditions, urban characteristics, and air pollution. The study focuses on New York City (NYC) as the study area, utilizing air pollutant readings collected from 155 monitoring stations within NYC over a period of 10 years (2009−2018). In addition to air quality data, meteorological conditions such as wind speed, temperature, precipitation, and urban parameters, including building area, traffic count, and land-cover data, are integrated into the analysis. The analysis considers urban parameters at different spatial scales, ranging from fine-scale resolution 25 × 25 m^2 cells to coarser resolutions of 5000 × 5000 m^2. The input to the prediction model consists of urban parameters at the selected spatial resolution, centered at the corresponding air monitoring station, and the meteorological conditions extracted from one weather station located in Central Park, NYC. We do not consider the urban properties in adjacent cells since they are explicitly included in the analysis at the next scale (for example, the adjacent cells in the 25 × 25 m^2 analysis, if important, will improve the prediction at the 50 × 50 m^2 scale). We also select only one meteorological station in the city to challenge the model under conditions of relative data scarcity, which are representative of most cities. This station thus represents synoptic scale dynamics, or the average over the city. Any departure of the finer conditions at the street canyon scale or at a specific location from that synoptic mean is assumed to be encoded in the urban data (building height, tree cover, etc.), and the ML model thus serves implicitly as a tool to downscale the meteorological data to the local scale.

2.1. Data Sets. *2.1.1. Air Quality Data.* The air pollution data set includes measurements for the period between 2009 and 2018 collected from 155 stations deployed by the New York City Community Air Survey (NYCCAS).[31−33] These data sets only provide the 2-week averages of the concentration of $PM_{2.5}$, NOx (NO and NO_2), SO_2, O_3, and EC. Figure 1 shows the locations of these stations. The effects of weather and urban parameters on air quality could vary with the season. For instance, in winter, the increased usage of heating systems in buildings can result in increased levels of indoor and outdoor air pollution, while in summer, increased solar radiation can lead to photochemical smog formation.[34] Therefore, seasonality has been recognized as an important factor in the context of analyzing the relationship between air quality and weather or urban parameters. As such, for each pollutant, we constructed three data sets representing the readings over the full period, readings over the winter cycle, and readings over the summer cycle. In this context, the winter

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

cycle represents the period between October and March, and the summer cycle covers the April−September period. The SO_2 and O_3 data were available only in the winter and summer, respectively.

The low temporal resolution of the data set (2-week averages) limits the possibility of addressing peak pollution at finer-than-seasonal scales, for example, for identifying days or periods of peak pollution, within the scope of this work. However, we are not aware of any publicly available data set that has the very fine spatial resolution and the range of pollutants of NYCCAS, as well as a finer temporal resolution. While predicting extreme events at shorter time scales remains of great interest, it would require not only data at finer temporal resolution but also different ML methods that may, for example, include time series forecasting methods such as long short-term memory neural networks to capture peak pollution periods. Future research could explore the integration of such time series forecasting techniques into our framework to predict extreme pollution events, as well as to explore the different drivers of air pollution across various time scales.

2.1.2. Urban Parameters. Since the spatial distribution of pollutants can vary greatly depending on the characteristics of the urban environment, a main aim here is to incorporate the variation in these characteristics and its influence on patterns of urban air quality. This will improve the prediction and enhance the analysis of such patterns. Since concentrations also vary significantly in time, we show in Table 1 the temporal and spatial standard deviations of each pollutant. Spatial variability can be just as significant as temporal variability in terms of the investigated pollutants within NYC, implying that a spatial and/or a temporal prediction machine learning model does not have a trivial task. This also highlights the importance of simultaneously considering the meteorology and the urban parameters, such as building area and characteristics and land use, when developing air quality prediction models.

Table 1. Temporal and Spatial Standard Deviation of the Air Pollutants, per Different Season Cycles[a]

pollutant	spatial standard deviation	temporal standard deviation
EC (absorbance)	0.243	0.393
EC_s (absorbance)	0.203	0.404
EC_w (absorbance)	0.261	0.383
NO_2(ppb)	5.039	6.223
NO_2_s (ppb)	3.436	6.592
NO_2_w (ppb)	4.435	5.845
NO (ppb)	11.26	11.523
NO_s (ppb)	6.083	9.695
NO_w (ppb)	9.661	13.391
O_3(ppb)	4.466	3.973
PM ($\mu g/m^3$)	2.793	1.937
PM_s ($\mu g/m^3$)	2.439	1.736
PM_w ($\mu g/m^3$)	2.998	2.137
SO_2(ppb)	1.914	1.189

[a]Subscript w for winter and s for summer, no subscript for full year data.

We use the data on the spatial characteristics of the urban environment for NYC curated by Llaguno-Munitxa and Bou-Zeid[16] to incorporate urban parameters. These parameters consist of building area, mean building heights, processed and interpolated traffic counts, and landcover data. The process, as detailed further by Llaguno-Munitxa and Bou-Zeid[16] involved several key steps. First, the values of the urban parameters were computed at various surrounding urban footprint scales, which ranged from 25 m × 25 m up to 5000 m × 5000 m. These scales allow one to capture the spatial variations in urban characteristics across the city. Next, statistical metrics for the selected urban parameters were derived on each of the footprint scales. These metrics included measures such as the building height, road area, tree canopy area, and green grass/shrub area. Furthermore, a traffic flow estimation model was employed, specific to the locations of the NYCCAS stations. This model accounted for the reported traffic counts at various locations within the city. The Primary Land Use Tax Lot Output (PLUTO)[35] data set was utilized to extract additional building-related metrics, including total building area, total residential and nonresidential area, and the mean building height. These metrics were computed for the NYCCAS air quality station locations and for each of the various footprint sizes considered.

The urban features considered in this work are listed in Table 2. Higher building densities and taller buildings can trap pollutants near the ground and reduce vertical mixing, leading to higher concentrations of different pollutants at the street level. Traffic emissions from cars, trucks, and buses release pollutants such as NOx and PM into the air, leading to a decrease in air quality. Land use can also impact air pollution, as areas with more vegetation can contribute to air pollution removal, while paved surfaces can lead to higher temperatures and increase the production of secondary pollutants. Note that, following the work in Llaguno-Munitxa and Bou-Zeid,[16] each of the considered features exists at 11 different spatial resolutions (centered at the location of the reading sensor), ranging from 25 m × 25 m to 5000 m × 5000 m resolution.

Table 2. Urban Parameters Used in this Study

parameter name	description	original data source
bldg. tot. area	total floor building areas	primary land use tax lot output[35]
bldg. height	mean of building height	[35]
nonres. tot. area	nonresidential total floor building area	[35]
res. tot. area	residential total floor building area	[35]
traffic	interpolated traffic count	NYCDOT[36]
trees	land use area −1: tree canopy	OpenData[37]
grass	land use area −2: grass/shrub	[37]
bare ear.	land use area −3: bare earth	[37]
water	land use area −4: water	[37]
bldg. grnd. area	land use area −5: buildings footprint area	[37]
roads	land use area− 6: roads	[37]
other paved	land use area− 7: other paved surfaces	[37]

2.1.3. Meteorological Data. We obtained the meteorological data from the US National Oceanic and Atmospheric Administration's (NOAA) Daily Summaries data set.[38] In particular, we collected the readings from the weather station USW00094728, located in New York City's Central Park. This station was deemed to best represent synoptic regional conditions with the least influence of the surrounding urban topology and characteristics. The data include daily measure-

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

ments of average wind speed (AWND), precipitation (PRCP), maximum and minimum temperatures (TMAX and TMIN), fastest 2 min wind speed (WSF2), and fastest 5 s wind speed (WSF5). We computed the summary statistics for each of these parameters, including mean, median, maximum, minimum, and standard deviation, for each 2-week period corresponding to the air quality readings, which provided a robust data set that accounts for the potential influence of meteorological factors on air pollution levels. We attempted to include other features such as wind direction, humidity, and cloud cover; however, such features did not affect predictive performance, probably due to the averaging over the 2-week periods.

2.2. Footprint Analysis. The analysis of the footprint scale of urban parameters is a crucial step in modeling and analyzing air quality in urban settings. Different urban features might have varying spatial footprints, meaning that their impact on pollutant concentrations may depend on the area of study. For instance, the effect of a nearby building on pollutant concentrations may be more significant at the local scale, say, within a few hundred meters, compared to the regional scale (several kilometers away). Furthermore, the effects of urban features on pollutant concentrations may vary, depending on the pollutant in question. For example, the impact of traffic-related emissions on the concentration of NO_2 is expected to be more pronounced at the local scale due to the short-lived nature of that pollutant. In contrast, the impact of urban vegetation on O_3 concentrations may be more significant at larger scales since the formation of O_3 requires time for the biogenic volatile organic compounds and NO_x to react and is influenced by regional-scale processes such as atmospheric transport.

Traditionally, one would opt to use the linear and multilinear regression models to determine and assess spatial variation of air pollution[39] and the correlation with urban features.[16] However, following this approach, one may not capture nonlinear relationships between the urban features and the different pollutants. The impact of a particular urban feature on pollutant concentrations may be more complex than what can be captured by a simple linear relationship. In this work, we leverage GBR to perform footprint analysis and highlight important scales. In this context, we train GBR using all available footprint sizes and extract the feature importance scores. These scores are used by GBR to assess the importance of each feature on the target variable; they are calculated based on how frequently each feature is used to make a split decision within the GBR-embedded decision trees (feature contribution to the reduction in the error). The use of GBR and feature importance analysis allowed us to capture complex, nonlinear relationships between urban features and air pollutants. This approach also provides us with a method to identify the optimal footprint sizes of different urban parameters and for each pollutant.

2.3. Gradient Boosting Regression. Gradient boosting regression (GBR)[40] is an ensemble-based machine learning algorithm that is often used for regression problems. GBR is composed of decision trees as base learners. Decision trees are versatile machine learning models commonly used for both classification and regression tasks. They work by recursively partitioning the data set based on the values of input features, leading to informed decisions and predictions. Conceptually, a decision tree resembles an inverted tree with a root node at the top and branching nodes that lead to terminal or leaf nodes. Each node represents a decision point where a specific feature is evaluated. This model operates by selecting the most informative feature and an associated threshold to split the data into two subsets. In this work, we used information gain as the criterion to determine the features and thresholds for splitting. GBR trains trees sequentially and focuses on reducing residual errors. This is in contrast to Random Forest, where trees are built in parallel, and the focus is on tree diversity.[41] In this study, GBR was utilized to model the relationship between air quality (represented by the concentration of several pollutants) and the various urban and weather parameters introduced before. GBR is a popular machine learning algorithm that is well-suited for predicting continuous variables and capturing nonlinear and complex relationships between input and output and is widely used for tackling environmental problems.[42−45] Moreover, GBR, similar to Random Forest (RF), provides feature importance scores, making it a popular choice for understanding the effects of various features on air quality. Feature importance provided by GBR can be used to understand the influence of each feature on the model's prediction. This is calculated based on how much a feature contributes to reducing the residual error in the model. Note that we used GBR for its ability to generate feature importance and since it displayed better performance relative to Random Forest and Linear Regression on our data.

3. RESULTS AND DISCUSSIONS

3.1. Footprint Scale of Urban Parameters. We train a GBR model for each air pollutant independently, at all available scales (25 × 25 to 5000 × 5000 m^2) of each urban parameter. We calculate the normalized relative feature importance scores by computing the total feature importance score for each pollutant and each urban parameter and across all scales and then dividing the feature importance of each parameter by the sum of all scores at the corresponding scale. These indicate the relative importance of each spatial scale of an urban parameter on the concentration of a given pollutant. We selected the scale with the highest feature importance score as the most informative. Initially, a unified scale for each urban parameter across all pollutants was used to streamline the approach for practitioners. This was done by averaging the feature importance of each scale and feature across pollutants and renormalizing the values to display them as relative percentages. The prediction process becomes simpler and more applicable to real-world scenarios when using unified scales, as it allows policymakers and urban planners to quickly identify effective strategies to mitigate air pollution in urban areas without the need for extensive calculations or data processing. Such an approach also ensures consistency in modeling the relationship between urban features and air pollutants across different pollutants, which can simplify the interpretation of the results and facilitate cross-pollutant comparisons. However, since different pollutants may have different spatial dependencies on urban features, using a unified local scale may not capture the full extent of the relationships between pollutants and urban parameters. To assess the predictability loss associated with this unified scale, we also tested the predictive model using different scales for different pollutants and confirmed that the use of a unified localized scale resulted in only a slight decrease in the accuracy of GBR, which was acceptable given the benefits of streamlining the prediction process. Table 3 shows the final unified spatial scales used for each urban parameter.

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

Table 3. Scales of the Urban Parameters Used in the Prediction, Which Featured the Highest Feature Importance Score, Unified Across all Pollutants

variable	scale (m × m)
bldg. tot. area	300 × 300
bldg. height	300 × 300
nonres. tot. area	500 × 500
res. tot. area	300 × 300
traffic	200 × 200
trees	500 × 500
grass	150 × 150
bare ear.	300 × 300
water	300 × 300
bldg. grnd. area	300 × 300
roads	500 × 500
other paved	100 × 100

The results of this footprint analysis are a refinement to the overall scales obtained previously by Llaguno-Munitxa and Bou-Zeid[16] but are broadly comparable. The scales produced herein provide valuable insights into the complex relationships between urban parameters and air quality at different scales, which can inform land-use planning, zoning, and traffic regulations, as well as the design and placement of green infrastructure (green roofs, trees, etc.).

3.2. Air Pollution Prediction. We evaluated the performance of GBR in predicting the concentration of air pollutants (independently) for various weather conditions and urban characteristics (with scales shown in Table 3), by following a train/test split of the data set, with a ratio of 80/20 for the training and testing sets, respectively. The train/test split is based on reading sites, ensuring that all measurements from a single site were in either the training or testing set. This step guarantees that the model is not biased toward a particular location, which would result in overfitting and inaccurate predictions for new and unseen data. In addition, a 10-fold cross-validation approach was employed to further reduce the risk of overfitting and to assess the generalizability of the model. In the context of the experiment, this cross-validation approach consisted of splitting the reading sites into 10 equal-sized subsets each containing 20% of the sites (i.e., 31 of the 155 stations, with overlaps between the subsets), choosing one subset as the testing group of sites, and training the model on all of the other sites. This process was repeated for each of the 10 subsets. We then use the average performance across the 10-fold to characterize the overall accuracy of the model in predicting the corresponding pollutant concentration. The computed standard deviation indicates the variability in the predictive performance across sites and the generalizability of the model. We adopted three evaluation metrics to assess the prediction performance of the model. These metrics include the percentage mean absolute error (PMAE), percentage root mean squared error (PRMSE), and the coefficient of determination (R^2). Note that PMAE and PRMSE are expressed as percentages to allow an easier comparison across different pollutants with different ranges of concentrations.

Table 4 shows the average results (and standard deviation) achieved by GBR. This trained and validated model can now be applied without the need for air quality data. An alternative approach would be to retrain a model to still use a handful (e.g., ≈ 5 sites) as inputs/features that give a representative mean across the city; it would be expected to give better prediction than the approach used here, but would require continuous monitoring at some sites.

Table 4. Average (and Standard Deviation) Performance of GBR Across the Pollutants Following the 10-Fold Cross-Validation Approach

pollutant	PMAE (%)	PRMSE (%)	R^2
PM	9.6 ± 0.8	15.0 ± 2.5	0.81 ± 0.04
PM_w	11.2 ± 1.4	16.7 ± 2.4	0.80 ± 0.04
PM_s	9.0 ± 1.9	13.1 ± 2.9	0.82 ± 0.06
EC	18.0 ± 2.1	26.1 ± 3.0	0.61 ± 0.07
EC_w	17.6 ± 1.8	25.7 ± 2.6	0.62 ± 0.06
EC_s	16.7 ± 2.4	24.7 ± 4.1	0.65 ± 0.12
NO	31.8 ± 3.7	47.0 ± 5.1	0.59 ± 0.09
NO_w	27.8 ± 6.1	38.7 ± 7.8	0.55 ± 0.13
NO_s	39.2 ± 4.8	61.6 ± 7.3	0.40 ± 0.18
NO_2	12.5 ± 1.3	17.7 ± 2.4	0.78 ± 0.07
NO_2_w	11.1 ± 1.4	15.6 ± 1.6	0.73 ± 0.05
NO_2_s	15.7 ± 2.1	21.3 ± 2.9	0.74 ± 0.08
SO_2_w	31.5 ± 3.3	45.8 ± 4.5	0.81 ± 0.03
O_3_s	8.1 ± 1.2	10.4 ± 1.4	0.72 ± 0.05

The best model performance is for PM, NO_2, and O_3; NO and SO_2 are clearly harder to predict. Based on the results of the 10-fold cross-validation, GBR achieved PMAE values between around 9% and 39%, with the lowest and highest values obtained for PM_s (summer cycle for $PM_{2.5}$) and NO_s (summer cycle for NO), respectively. The average PRMSE values ranged from around 13% to 62%, with the lowest and highest values obtained as well for PM_s and NO_s, respectively. The average R^2 values ranged from 0.4 to 0.8, with the highest and lowest values obtained for SO_2 and NO_s, respectively. The GBR model showed good performance across most pollutants. However, there were some variations in the performance for different pollutants, which could be attributed to the underlying characteristics of each pollutant and the availability and quality of the data. For instance, $PM_{2.5}$ (both winter and summer) showed high R^2 values >0.8, indicating that the models were able to explain a large proportion of the variability in this pollutant. NO (both winter and summer) showed relatively lower R^2 values ranging from 0.4 to 0.6. This could be because NO has a more complex relation with urban parameters and more active chemistry (e.g., shade of trees or buildings impacting on photochemistry), which may not be fully captured by the framework used in this study. We further employed GBR to predict NO_x ($NO + NO_2$) and found an R^2 of 0.7 (intermediate between NO and NO_2 performance). This shows that the uncertainty in predicting NO is not related to the oxidation of NO to NO_2.

Table 4 shows that slight variations were observed in the performance of GBR between the winter and summer cycles for the same pollutant (2−3% error difference), except for NO where the seasonal differences are quite significant (the error was lower in winter than in summer). This suggests that seasonal variations in meteorological conditions might in general have a small impact on the performance of the model. In addition, wind direction, relative humidity, and cloud cover did not increase the predictive performance of the model. This is probably due to the averaging over 2-week periods, which renders the model incapable of fully capturing the variability of these features.

In Figure 2, the scatter plots between the observed and predicted (GBR) concentration values of the different

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

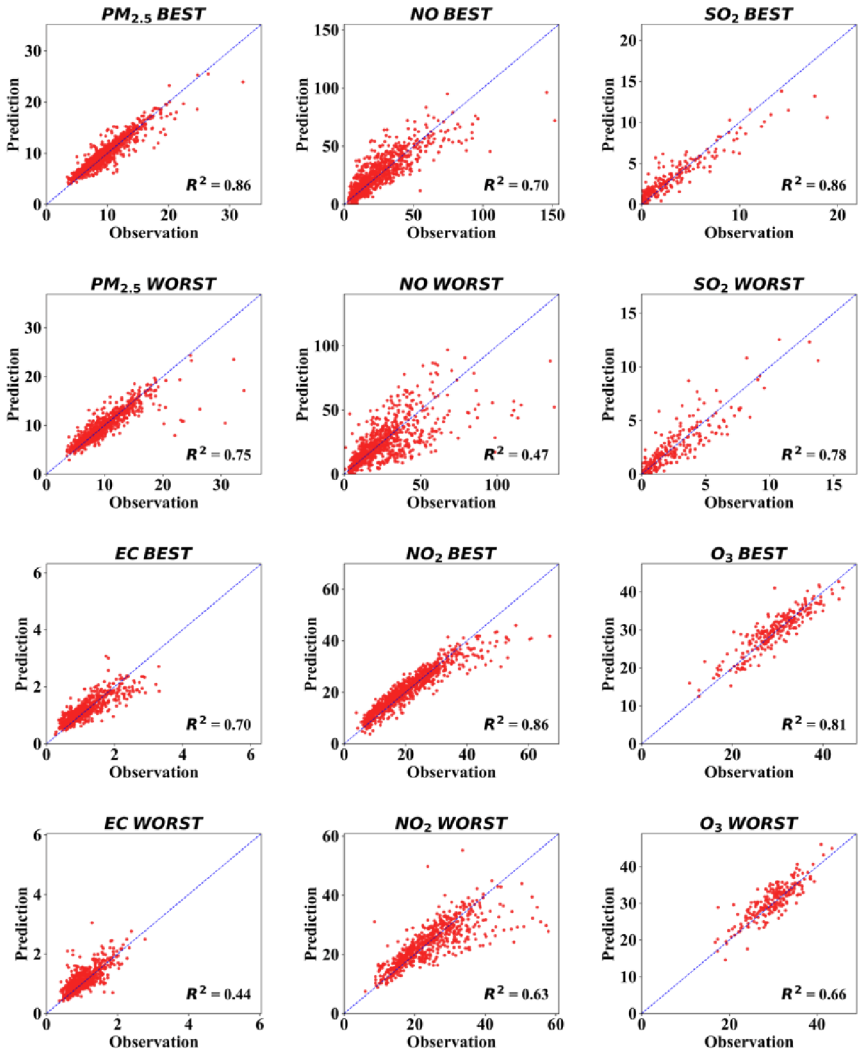


Figure 2. Scatter plots showing observed versus predicted (by GBR) concentration values of each pollutant based on worst (min) and best (max) fold in terms of R^2 across the 10 folds. The diagonal dashed line represents a perfect match between the observed and predicted values.

pollutants are shown. The best and worst cases (in terms of R^2) scatter plots across the 10-folds for each pollutant are displayed, to showcase the difference in performance when different subsets of sites are selected as training/testing. These plots can provide insight into the characteristics of the data, such as the range of concentration values and the distribution of the residuals. The range from the minimum to the maximum R^2 values is fairly large across most pollutants, indicating that the performance of GBR can vary appreciably depending on the split of training and testing sites, reflecting the importance of cross-validation. For instance, R^2 for EC varied from around 0.44 to 0.7, while for PM they ranged from around 0.75 to 0.86. Although such variation might also depend on the choice of the model and its hyperparameter, the impact of these factors is less pronounced than that of the monitoring sites, as shown by the relatively consistent performance of the GBR across different pollutants. Overall, the range of R^2 values across the 10 folds underscores the importance of cross validation and the potential variability of ML model skills depending on the sites used for training, and on the location the model is being applied. The optimal selection of urban observational sites for training is an important theme itself.[46,47] While we will not delve into such a selection process in this paper, in the next section, the experiment will be further expanded to perform error-based site-specific analysis and thus understand the impact of testing sites selection.

3.3. Mapping Site-Specific Errors. Figure 2, and the variability between minimum and maximum R^2 values for the same pollutant, highlight the differences in GBR predictive performance based on the training/testing sites selection. This finding motivated us to perform a more detailed analysis to highlight the predictive performance at each site, identify specific sites where the predictive model may be underperforming, and uncover particular features that are affecting such a lack of predictive skill. An error-based site-specific analysis was performed in which the prediction experiment using GBR with 500-fold cross-validation was repeated. The usage of a relatively high number of folds (500) was motivated by several factors. First, we aimed to perform a detailed site-specific error analysis to identify potential underperforming sites and understand the predictive performance variations across monitoring locations. Second, by using a large number of folds, we could ensure that each monitoring site appeared in the testing set in at least 100 experiments. This was crucial for the calculation of the prediction error at each site to converge to its statistical expected value, as it allowed for a robust

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

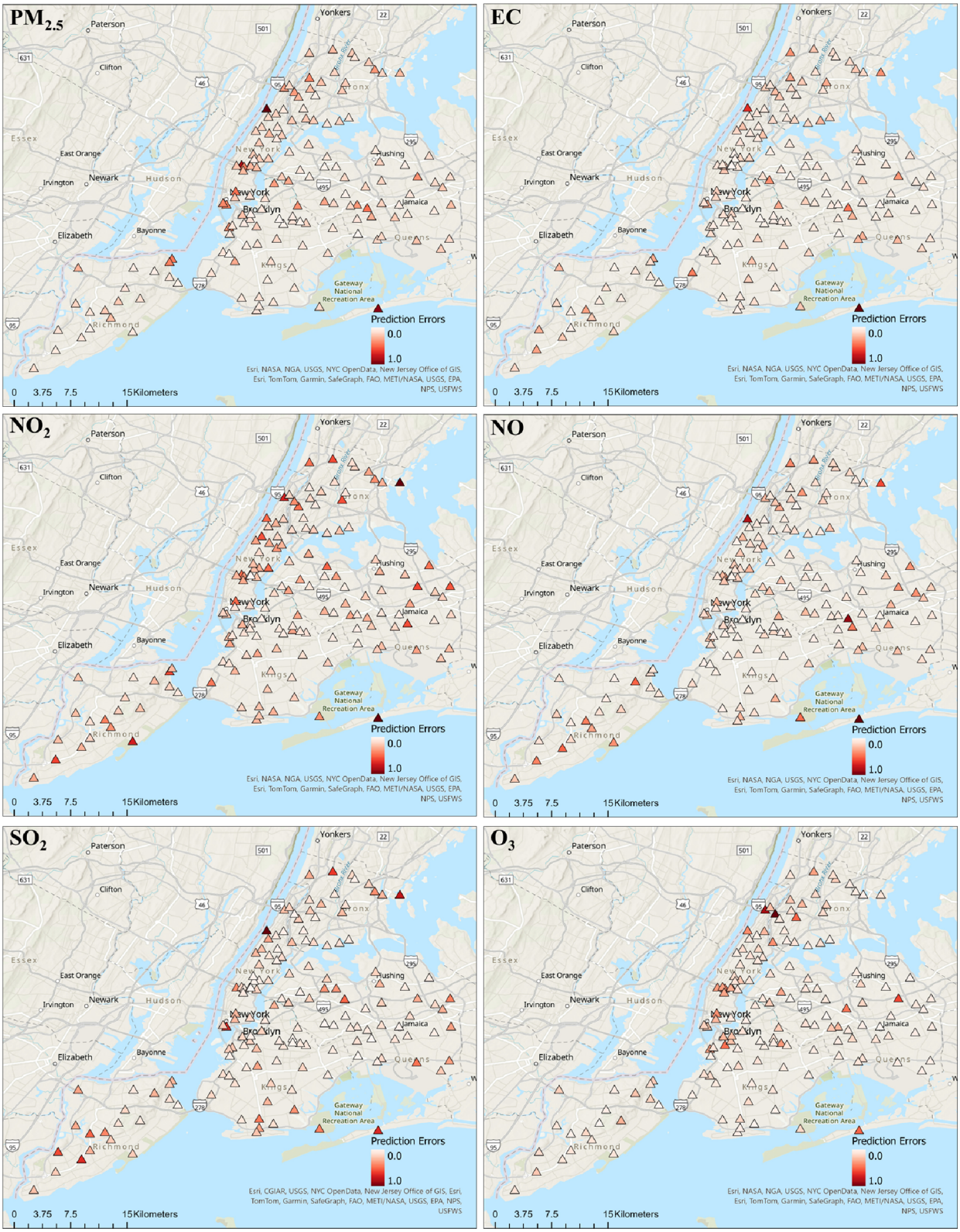


Figure 3. Maps showing the spatial distribution of site-based MAE values for the air pollutants (PM2.5, EC, NO, NO_2, SO_2, and O_3) in the study area. Higher MAE values are indicated by warmer colors, ranging from light (low MAE) to dark (high MAE) red. The maps reveal spatial patterns and hotspots of PMAE values for each pollutant, highlighting areas where air pollution predictions may be less accurate. Maps created by the authors using the ArcGIS Pro software.

assessment of the model's performance across diverse urban areas. The MAE per site for each pollutant was calculated, averaging across the predictions of 500 testing experiments.

Figure 3 reports the maps of the normalized MAE (between 0 and 1) over NYC for each pollutant. Each circle denotes a monitoring site with colors ranging from light (low error) to

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

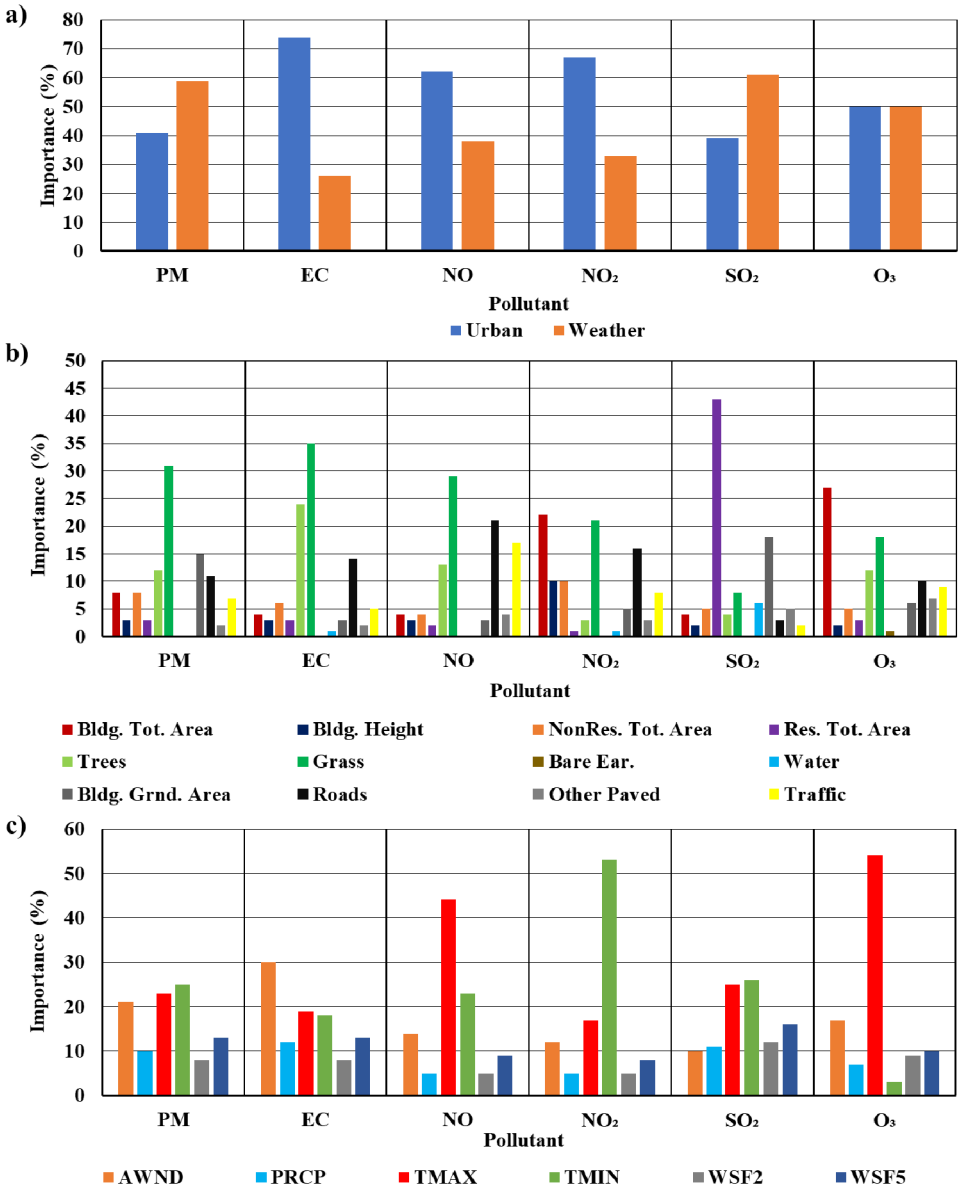


Figure 4. Feature importance a) categorized into urban and meteorological features, b) of urban parameters, and c) of meteorological parameters for each pollutant. The bar charts show the percentage of importance for each category.

dark (high error) red. The consistent locations with high error across pollutants, mostly located at the western edges of the station coverage map and near two main NYC entry points (the Lincoln Tunnel and the George Washington Bridge, both connecting Manhattan to New Jersey), could indicate that the error is highly influenced by localized high traffic that might not be captured by the traffic data used here. The fact that these high-error locations are consistent across multiple pollutants suggests that there may be some common underlying factors contributing to the error. Additional analysis is needed to confirm these hypotheses and investigate other potential factors that may contribute to the observed patterns. The results, however, confirm the importance of site-specific analysis, as the performance of the GBR, and machine learning more broadly, will vary significantly across monitoring sites, suggesting that some sites are dominated by hyperlocal drivers.

3.4. Importance of Urban and Meteorological Features. GBR enables us to compute feature importance scores, which indicate the relative influence of each feature on the prediction of the target variable. A feature score is expressed as a percentage of the total importance of all features. The input features used in the study were categorized into urban (listed in Table 2) and weather features.

Figure 4a shows the total feature importance summed per category (urban and weather) for each pollutant. The importance of the input features for the two categories varies across pollutants. In general, urban features were more important for EC, NO, and NO_2, while meteorological features were more important than urban features for SO_2 and $PM_{2.5}$. For O_3, both categories had similar importance. These results suggest that for pollutants strongly linked to vehicular emissions (EC, NO, and NO_2) urban characteristics, such as traffic count and building area and type, can be strong predictors of air pollution levels. While for pollutants that have secondary formation pathways (PM2.5) or nontraffic sources (SO_2, which is emitted from point and fixed combustion sources), meteorological conditions, such as wind speed and temperature, are more influential.

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

These results highlight the importance of carefully considering the choice of input features and their relative importance when developing predictive frameworks for specific species. Note that the number of features within each category did not have an effect on the feature importance results.

Figure 4b shows the normalized feature importance of urban parameters. In general, building areas, nonresidential areas, and vegetated cover (trees and grass/shrubs) showed most the influence on predicting air quality. However, the importance of these features differed significantly among pollutants. For example, the building area was most important for NO_2 and O_3, while grass cover was most important for PM and EC. The residential total area was relatively unimportant in predicting most pollutants except for SO_2 (emitted in NYC primarily from home heating systems) where it had the highest feature importance. These findings suggest that the influence of different urban parameters on air quality varies greatly depending on the pollutant being considered, again highlighting the importance of pollutant-specific modeling approaches. One should be made of the low predictive performance of GBR on NO concentration, resulting in low confidence in the feature importance results for that pollutant.

Figure 4c shows the feature importance of the weather parameters. Note that for each weather parameter, the importance of all corresponding feature (mean, median, max, min, and standard deviation) were combined into one value. However, daily TMAX and TMIN are treated as different weather parameters, each with its own statistics; therefore, the feature importance of the mean, median, max, and min for each of them (daily maximum and daily minimum) are combined separately. The importance of weather parameters varied across different pollutants as well. For instance, TMAX (maximum temperature) has the highest importance for NO and O_3, while it has a relatively lower importance for NO_2, for which TMIN is the most important. The increased sensitivity of NO_2 to TMIN can be attributed to several factors. One key aspect is that NO_2 levels tend to peak prior to the onset of morning emissions, a period associated with the absence of effective boundary layer transport and mixing due to the preceding stable atmospheric stratification during the night. A lower TMIN under colder conditions increases the likelihood of stronger temperature inversions, which can lead to reduced vertical mixing of air pollutants and higher NO_2 concentrations. This limited mixing capacity creates conditions conducive to the accumulation and buildup of NO_2, as emissions are trapped near the surface.[48] The behavior of NO_2 in our data set agrees with such occurrence, showing a negative correlation with TMIN (with a Pearson coefficient of −0.65). It is important to note that while these observations provide valuable insights, a comprehensive understanding of the underlying causes would necessitate the use of detailed photochemical, aerosol, and transport coupled models. Furthermore, AWND (average wind speed) and WSF5 (fastest 5 s wind speed) seemed to have relatively consistent importance across different pollutants. PRCP (precipitation), on the other hand, has relatively low importance across all pollutants except for PM and SO_2, which highlights the importance of wet deposition removal mechanisms for these but not the other pollutants. These results underscore the complex impact of meteorological conditions on air quality and why understanding these relationships is crucial for developing effective air quality management strategies. These results also confirm the unique properties and dynamics of air pollutants. For instance, PM should have a dependence upon precipitation, as the dominant loss term for aerosol particles is rainout. The measurements of TMAX are highly relevant for ozone, as warmer/higher temperatures, coupled with potentially sunnier conditions, contribute to increased ozone production.

Overall, attribution analysis shows that pollutants dominated by urban characteristics (NO_x and EC) can be effectively mitigated by urban planning and design and that the scale of intervention should be 500 × 500 m^2 or larger for a significant improvement to be realized. Pollutants that are mostly controlled by weather variability (PM and SO_2, and to some extent O_3) cannot be attenuated much by modifying urban form and function, but only through source reduction. The intricate interaction of weather and urban drivers will surely persist at smaller time scales that our data set cannot capture, but the dynamics may be even more complex. Data sets with finer temporal resolutions are needed to probe, for example, the impact of transient weather events or specific conditions such as rainy days, heatwaves, or windy periods.

The findings have several implications for air quality management, policy, and practice within urban settings. The significant drivers of air pollution in urban areas and their spatial scale can inform the development of targeted policies and interventions to enhance air quality and the spatial scale at which these interventions need to be deployed. For example, elevated buildings and low green cover were clearly associated with adverse air quality, and thus, increasing green cover and lowering building height should improve air quality. But here one has to be mindful that attribution analysis can only, in principle, indicate that green cover was associated with improved air quality; it cannot fully elucidate or confirm causality. For example, this association can be due to lower traffic in greener neighborhoods or more spacing between buildings that promotes ventilation rather than to any pollution abatement properties of vegetation. Mitigation decisions can thus be guided by attribution analysis but should also be informed by physical processes that associate a driver with the output it influences. Our findings can however confirm that such dense neighborhoods with lower green cover are particularly vulnerable, and thus emission controls in these neighborhoods should be tighter (e.g., converting them to low emission zones).

Our spatial predictive framework, while providing valuable insights into air quality patterns, does come with certain limitations and uncertainties that warrant consideration. First, our analysis operates at a spatial resolution dictated by the deployed monitoring network, potentially missing finer-scale variations in air quality. Mobile air quality data that are now increasingly available may allow future applications at finer scales.[49] Furthermore, the model's performance is influenced by data availability, and it would surely require retraining for application to other cities, as well as urban data for those cities, which may not be readily available. Efforts to standardize and improve data collection practices can enhance the applicability of ML models in general and the present one in particular. While these limitations are important to consider, our model and analyses remain valuable for understanding urban air quality drivers broadly as well for spatial prediction in NYC. Furthermore, our study develops easily replicable and transferable methods. This transferability, however, requires urban parameters that are extensive and expensive to measure and extract for urban areas. This will challenge our model and

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

similar ML approaches for many cities, limiting the applicability of the framework. Simpler approaches for "characterizing the city" are needed, such as using only satellite images to capture urban parameters that are often hard to collect.

The city of New York served as a compelling demonstration site due to its diverse urban characteristics, extensive monitoring infrastructure, and well-documented air quality challenges. However, these very qualities further complicate the transferability of our approach to other urban areas worldwide. While cities across the globe grapple with similar air quality concerns, the intensity and sources of pollution vary considerably. This would hence require tailoring of any ML framework to accommodate the unique urban features and weather patterns of different cities. Nevertheless, the importance of considering the interplay between urban characteristics and meteorological factors in assessing air quality can be seen as a universal concept that transcends geographical boundaries.

3.5. Temporal Forecasting of Air Quality. Thus far, this study has been strictly focused on spatial prediction at "unseen" locations. That is, given air quality data at a given time from a network of sensors, we were able to predict the concurrent concentrations at all other locations where no observations were made in the city. The trained model does not even need the observations as input features, since they were encoded during the training into the GBR model. Application of such spatial interpolation models would be to obtain complete maps of air quality over a city from a discrete set of observations.

In this section, we extend our analysis to address an equally important aspect: the temporal dynamics of the pollutant concentrations. Capturing these temporal patterns in pollutant concentrations is crucial for the development of accurate predictive models that are needed for short-term mitigation and risk evaluation by city managers. While the spatial analysis in previous sections established that the dynamics of air pollutants are influenced by meteorological conditions, urban features, and localized sources, the work herein adds a temporal dimension to our predictions. In addition to the previously used input features, we incorporate past measurements from a single location or station in the city (not the same as the location being predicted). This reference station serves as a representative point for capturing the mean temporal change in pollutant dynamics over the city. By leveraging this information, we extend our predictions to other sites throughout New York City, thereby forecasting the temporal changes in pollutant concentrations across the urban landscape.

To perform this temporal forecasting of each pollutant concentration, we adopt a time-based train-test split approach instead of the site-based split used in the previous experiment. We then selected the single reference site and incorporated its reading as an input feature within our predictive model. This mimics a scenario where a dense network is deployed in the city for a limited period of time to train the GBR algorithm and then that network is moved or removed, leaving only a single reference station.

For this experiment, we chose the site with the least missing values, which was located in Central Park, as the reference station. This does not guarantee that the site is necessarily the most informative but mimics the case where other considerations, besides the representativeness of the station's data, dictate which location is maintained. This reference site then serves as a representative location for capturing temporal patterns in pollutant concentrations. Table 5 shows the results achieved by GBR when training on the period between 2008 and 2015 and testing on the 2016−2018 period.

Table 5. Performance of GBR Across the Pollutants for the Temporal Mapping of Spatial Variability

pollutant	PMAE (%)	PRMSE (%)	R^2
PM	13.70	19.28	0.65
EC	20.36	27.73	0.63
NO	34.01	45.86	0.63
NO_2	10.78	16.00	0.82
SO_2	182.51	302.19	
O_3	6.69	8.97	0.69

Overall, GBR showed varying performance across different pollutants. For EC and NO, the model achieved moderate accuracy in terms of PMAE and reasonable R^2 scores. For NO_2, PM, and O_3, GBR performed relatively well, with lower PMAE values and higher R^2 scores compared with EC and NO, indicating that GBR was able to capture a significant portion of the variability in these pollutant concentrations and provide accurate predictions. It is important to note that the negative R^2 value obtained for SO_2 indicates a poor fit of the model to the data, suggesting that the GBR model cannot capture the underlying patterns and relationships for SO_2. The most likely explanation of this poor performance is that the chosen reference site is not in a residential neighborhood where SO_2 emissions are concentrated and is thus a poorly representative surrogate for past concentration. This could indicate that SO_2 concentration evolution through time possesses a relatively higher spatial heterogeneity (as in fact indicated by the ratio of spatial standard deviation to temporal standard deviation for SO_2 in Table 4, which is the highest among the pollutant considered here). The motivation behind using a reference site is to capture the mean temporal change in pollutant dynamics over the city, while the GBR model captures spatial variability.

Cities can also exploit their sensitivity to weather for better forecasting and warning. On that note, machine learning approaches can be highly effective in augmenting air quality information in cities. The two illustrations we produced in this study pertain to using GBR models to (i) predict air pollutants at different locations from concurrent discrete sensing locations (spatial interpolation at unseen locations) and (ii) forecast pollutant concentrations for the city, after an initial deployment of a dense network for training, by only retaining a few (here one) reference stations and leveraging urban and weather information (temporal forecasting to unseen times). While the data we have here did not allow forecasting at fine temporal resolutions for examining short-term events (due to the two-week averaging), the results indicate that GBR is likely to be effective for such applications and can be combined with advanced time-series modeling approaches as discussed previously, which would allow cities to inform their populations of hot spots and hot periods of air pollution risk and to take actions based on very fine spatial and temporal air quality information.

https://doi.org/10.1021/acs.est.4c00783
Environ. Sci. Technol. 2024, 58, 6313−6325

■ ASSOCIATED CONTENT

Data Availability Statement

Pollutant concentrations are collected from New York City Community Air Survey (NYCCAS, https://www.nyc.gov/site/doh/data/data-sets/air-quality-nyc-community-air-survey.page). Urban data are collected from PLUTO (https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page), provided by the NYC Department of City Planning. Traffic data are collected from the traffic volume count database from NYC open data server (https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts/btm5-ppia). Meteorological data are collected from the US National Oceanic and Atmospheric Administration's (NOAA) Daily Summaries for the weather station USW00094728 (https://www.ncei.noaa.gov/maps/daily-summaries/).

■ AUTHOR INFORMATION

Corresponding Author

Elie Bou-Zeid − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States*; Email: ebouzeid@princeton.edu

Authors

Firas Gerges − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States*; orcid.org/0000-0003-0640-9782

Maider Llaguno-Munitxa − *Louvain Research Institute of Landscape, Architecture, Built Environment, UCLouvain, Ottignies-Louvain-la-Neuve 1348, Belgium*

Mark A. Zondlo − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States*; orcid.org/0000-0003-2302-9554

Michel C. Boufadel − *Center for Natural Resources, Department of Civil and Environmental Engineering, New Jersey Institute of Technology, University Heights, Newark, New Jersey 07102, United States*; orcid.org/0000-0002-5994-663X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c00783

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The study was supported by the Samsung Advanced Institute of Technology and the Army Research Office under contract W911NF2010216.

■ REFERENCES

(1) Kampa, M.; Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **2008**, *151*, 362−367.

(2) Mabahwi, N. A. B.; Leh, O. L. H.; Omar, D. Human health and wellbeing: Human health effect of air pollution. *Procedia Soc. Behav. Sci.* **2014**, *153*, 221−229.

(3) Van Kamp, I.; Leidelmeijer, K.; Marsman, G.; De Hollander, A. Urban environmental quality and human well-being: Towards a conceptual framework and demarcation of concepts; a literature study. *Landscape Urban Plann.* **2003**, *65*, 5−18.

(4) McDuffie, E.; Martin, R.; Yin, H.; Brauer, M. Global burden of disease from major air pollution sources (GBD MAPS): A global approach; Research Reports: Health Effects Institute, 2021

(5) Vohra, K.; Marais, E. A.; Bloss, W. J.; Schwartz, J.; Mickley, L. J.; Van Damme, M.; Clarisse, L.; Coheur, P.-F. Coheur, Rapid rise in premature mortality due to anthropogenic air pollution in fast-growing tropical cities from 2005 to 2018. *Sci. Adv.* **2022**, *8*, No. eabm4435.

(6) Weichenthal, S.; Hatzopoulou, M.; Goldberg, M. S. Exposure to traffic-related air pollution during physical activity and acute changes in blood pressure, autonomic and micro-vascular function in women: A cross-over study. *Part. Fibre Toxicol.* **2014**, *11*, 1−16.

(7) WHO *Ambient air pollution: A global assessment of exposure and burden of disease*; World Health Organization, 2016

(8) Borck, R.; Schrauth, P. Population density and urban air quality. *Reg. Sci. Urban Econ.* **2021**, *86*, 103596.

(9) Chen, J.; Wang, B.; Huang, S.; Song, M. The influence of increased population density in China on air pollution. *Sci. Total Environ.* **2020**, *735*, 139456.

(10) Nagendra, S. S.; Venugopal, K.; Jones, S. L. Assessment of air quality near traffic intersections in Bangalore city using air quality indices. *Transp. Res. Part D: Transp. Environ.* **2007**, *12*, 167−176.

(11) Comert, G.; Darko, S.; Huynh, N.; Elijah, B.; Eloise, Q. Evaluating the impact of traffic volume on air quality in South Carolina. *Int. J. Transp. Sci. Technol.* **2020**, *9*, 29−41.

(12) Hatzopoulou, M.; Miller, E. J. Linking an activity-based travel demand model with traffic emission and dispersion models: Transport's contribution to air pollution in Toronto. *Transp. Res. Part D: Transp. Environ.* **2010**, *15*, 315−325.

(13) Hankey, S.; Marshall, J. D. Urban form, air pollution, and health. *Curr. Environ. Health Rep.* **2017**, *4*, 491−503.

(14) Kang, J. E.; Yoon, D.; Bae, H.-J. Evaluating the effect of compact urban form on air quality in Korea. *Environ. Plann. B: Urban Anal. City Sci.* **2019**, *46*, 179−200.

(15) Iodice, P.; Senatore, A. Industrial and urban sources in Campania, Italy: The air pollution emission inventory. *Energy Environ.* **2015**, *26*, 1305−1317.

(16) Llaguno-Munitxa, M.; Bou-Zeid, E. The environmental neighborhoods of cities and their spatial extent. *Environ. Res. Lett.* **2020**, *15*, 074034.

(17) Li, C.; Wang, Z.; Li, B.; Peng, Z.-R.; Fu, Q. Investigating the relationship between air pollution variation and urban form. *Build. Environ.* **2019**, *147*, 559−568.

(18) Yuan, C.; Ng, E.; Norford, L. K. Improving air quality in high-density cities by understanding the relationship between air pollutant dispersion and urban morphologies. *Build. Environ.* **2014**, *71*, 245−258.

(19) Alpaydin, E.. *Introduction to machine learning*; MIT press, 2020.

(20) Zhao, B.; Yu, L.; Wang, C.; Zhu, J.; Qu, S.; Taiebat, M. Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning. *Environ. Sci. Technol.* **2021**, *55* (8), 5579−5588.

(21) Yao, J.; Brauer, M.; Raffuse, S.; Henderson, S. B. Machine learning approach to estimate hourly exposure to fine particulate matter for urban, rural, and remote populations during wildfire seasons. *Environ. Sci. Technol* **2018**, *52*, 13239−13249.

(22) Ren, X.; Mi, Z.; Cai, T.; Nolte, C. G.; Georgopoulos, P. G. Flexible Bayesian ensemble machine learning framework for predicting local ozone concentrations. *Environ. Sci. Technol* **2022**, *56*, 3871−3883.

(23) Song, J.; Fan, H.; Gao, M.; Xu, Y.; Ran, M.; Liu, X.; Guo, Y. Toward high-performance map-recovery of air pollution using machine learning. *ACS ES&T Eng.* **2023**, *3*, 73−85.

(24) Song, J.; Han, K.; Stettler, M. E. Deep-MAPS: Machine-learning-based mobile air pollution sensing. *IEEE Internet Things J.* **2021**, *8*, 7649−7660.

(25) Chai, J.; Song, J.; Zhang, L.; Guo, B.; Xu, Y.; Li, Y. Optimization of Land Use Regression Modelling of PM 2.5 Spatial Variations in Different Seasons across China. *J. Sens.* **2022**, *2022*, 1−9.

(26) Li, T.; Zhang, Q.; Peng, Y.; Guan, X.; Li, L.; Mu, J.; Wang, X.; Yin, X.; Wang, Q. Contributions of various driving factors to air pollution events: Interpretability analysis from Machine learning perspective. *Environ. Int.* **2023**, *173*, 107861.

(27) Zhang, B.; Zhang, Y.; Zhang, K.; Zhang, Y.; Ji, Y.; Zhu, B.; Liang, Z.; Wang, H.; Ge, X. Machine learning assesses drivers of PM2.

5 air pollution trend in the Tibetan Plateau from 2015 to 2022. *Sci. Total Environ.* **2023**, *878*, 163189.
(28) Qi, M.; Dixit, K.; Marshall, J. D.; Zhang, W.; Hankey, S. National land use regression model for no2 using street view imagery and satellite observations. *Environ. Sci. Technol.* **2022**, *56*, 13499−13509.
(29) Hankey, S.; Marshall, J. D. Land use regression models of on-road particulate air pollution (particle number, black carbon, PM2. 5, particle size) using mobile monitoring. *Environ. Sci. Technol.* **2015**, *49*, 9194−9202.
(30) Azmi, W. N. F. W.; Pillai, T. R.; Latif, M. T.; Koshy, S.; Shaharudin, R. Application of Land Use Regression Model to Assess Outdoor Air Pollution Exposure: A Review. *Environ. Adv.* **2023**, *11*, 100353.
(31) Matte, T. D.; Ross, Z.; Kheirbek, I.; Eisl, H.; Johnson, S.; Gorczynski, J. E.; Kass, D.; Markowitz, S.; Pezeshki, G.; Clougherty, J. E. Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York City: design and implementation. *J. Exposure Sci. Environ. Epidemiol.* **2013**, *23*, 223−231.
(32) Pitiranggon, M.; Johnson, S.; Haney, J.; Eisl, H.; Ito, K. Long-term trends in local and transported PM2. 5 pollution in New York City. *Atmos. Environ.* **2021**, *248*, 118238.
(33) Johnson, S.; Bobb, J. F.; Ito, K.; Savitz, D. A.; Elston, B.; Shmool, J. L.; Dominici, F.; Ross, Z.; Clougherty, J. E.; Matte, T. Ambient fine particulate matter, nitrogen dioxide, and preterm birth in New York City. *Environ. Health Perspect.* **2016**, *124*, 1283−1290.
(34) Dickerson, R.; Kondragunta, S.; Stenchikov, G.; Civerolo, K. L.; Doddridge, B.; Holben, B. N. The impact of aerosols on solar ultraviolet radiation and photochemical smog. *Science* **1997**, *278*, 827−830.
(35) Pluto. *PLUTO and MapPLUTO*. https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page; NYC Department of City Planning. (accessed 2018−11−01).
(36) NYCDOT. *Traffic Volume Counts (2012−2013)*. https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts/btm5-ppia; City of New York. (accessed 2018−11−01).
(37) N. OpenData. *Landcover Raster Data (2010) − 3 ft Resolution*. https://data.cityofnewyork.us/widgets/9auy-76zt; City of New York. (accessed 2018−11−01).
(38) NOAA. *Daily Summaries Station Details*. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail. (accessed 2022−12−22).
(39) 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.
(40) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29*, 1189−1232.
(41) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(42) Chen, Z.-Y.; Zhang, T.-H.; Zhang, R.; Zhu, Z.-M.; Yang, J.; Chen, P.-Y.; Ou, C.-Q.; Guo, Y. Extreme gradient boosting model to estimate PM2. 5 concentrations with missing-filled satellite data in China. *Atmos. Environ.* **2019**, *202*, 180−189.
(43) Su, Y. *Prediction of air quality based on Gradient Boosting Machine Method*; IEEE, 2020.
(44) Cai, J.; Xu, K.; Zhu, Y.; Hu, F.; Li, L. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Appl. Energy* **2020**, *262*, 114566.
(45) Cui, Z.; Qing, X.; Chai, H.; Yang, S.; Zhu, Y.; Wang, F. Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis. *J. Hydrol.* **2021**, *603*, 127124.
(46) Malings, C.; Pozzi, M.; Klima, K.; Bergés, M.; Bou-Zeid, E.; Ramamurthy, P. Surface heat assessment for developed environments: Optimizing urban temperature monitoring. *Build. Environ.* **2018**, *141*, 143−154.
(47) Yang, J.; Bou-Zeid, E. Designing sensor networks to resolve spatio-temporal urban temperature variations: fixed, mobile or hybrid? *Environ. Res. Lett.* **2019**, *14*, 074022.
(48) Li, J.; Wang, Y.; Zhang, R.; Smeltzer, C.; Weinheimer, A.; Herman, J.; Boersma, K. F.; Celarier, E. A.; Long, R. W.; Szykman, J. J.; et al.et al. Comprehensive evaluations of diurnal NO_2 measurements during DISCOVER-AQ 2011: effects of resolution-dependent representation of NO_x emissions. *Atmos. Chem. Phys.* **2021**, *21*, 11133−11160.
(49) Llaguno-Munitxa, M.; Bou-Zeid, E. In Sensing the environmental neighborhoods. In *Proceedings of the 2020 Digital FUTURES. CDRF 2020*; Yuan, P. F.; Yao, J.; Yan, C.; Wang, X.; Leach, N., Eds.; Springer: Singapore, 2021. DOI: 10.1007/978-981-33-4400-6_12