ARTICLE

Non-linear probabilistic calibration of low-cost environmental air pollution sensor networks for neighborhood level spatiotemporal exposure assessment

Andrew Patton[1,2✉], Abhirup Datta[3], Misti Levy Zamora[1,4,5], Colby Buehler[5,6], Fulizi Xiong[7], Drew R. Gentner[5,6] and Kirsten Koehler[1,4]

BACKGROUND: Low-cost sensor networks for monitoring air pollution are an effective tool for expanding spatial resolution beyond the capabilities of existing state and federal reference monitoring stations. However, low-cost sensor data commonly exhibit non-linear biases with respect to environmental conditions that cannot be captured by linear models, therefore requiring extensive lab calibration. Further, these calibration models traditionally produce point estimates or uniform variance predictions which limits their downstream in exposure assessment.
OBJECTIVE: Build direct field-calibration models using probabilistic gradient boosted decision trees (GBDT) that eliminate the need for resource-intensive lab calibration and that can be used to conduct probabilistic exposure assessments on the neighborhood level.
METHODS: Using data from Plantower A003 particulate matter (PM) sensors deployed in Baltimore, MD from November 2018 through November 2019, a fully probabilistic NGBoost GBDT was trained on raw data from sensors co-located with a federal reference monitoring station and compared against linear regression trained on lab calibrated sensor data. The NGBoost predictions were then used in a Monte Carlo interpolation process to generate high spatial resolution probabilistic exposure gradients across Baltimore.
RESULTS: We demonstrate that direct field-calibration of the raw $PM_{2.5}$ sensor data using a probabilistic GBDT has improved point and distribution accuracies compared to the linear model, particularly at reference measurements exceeding 25 $\mu g/m^3$, and also on monitors not included in the training set.
SIGNIFICANCE: We provide a framework for utilizing the GBDT to conduct probabilistic spatial assessments of human exposure with inverse distance weighting that predicts the probability of a given location exceeding an exposure threshold and provides percentiles of exposure. These probabilistic spatial exposure assessments can be scaled by time and space with minimal modifications. Here, we used the probabilistic exposure assessment methodology to create high quality spatial-temporal $PM_{2.5}$ maps on the neighborhood-scale in Baltimore, MD.

IMPACT STATEMENT

- We demonstrate how the use of open-source probabilistic machine learning models for in-place sensor calibration outperforms traditional linear models and does not require an initial laboratory calibration step. Further, these probabilistic models can create uniquely probabilistic spatial exposure assessments following a Monte Carlo interpolation process.

Keywords: Exposure modeling; Air pollution; Sensors; Geospatial analyses

Journal of Exposure Science & Environmental Epidemiology (2022) 32:908–916; https://doi.org/10.1038/s41370-022-00493-y

INTRODUCTION

According to The World Health Organization (WHO), fine particulate matter ($PM_{2.5}$) is responsible for approximately 7 million premature mortalities per year [1]. Within the United States, 88,000 annual deaths are attributed to $PM_{2.5}$ exposure [2]. Further, $PM_{2.5}$ is considered a Group 1 carcinogen according to the

[1]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, 615N. Wolfe St., Baltimore, MD 21205, USA. [2]Geospatial Analysis Lab, Harney Science Center, University of San Francisco, San Francisco, CA 94117, USA. [3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615N. Wolfe St., Baltimore, MD 21205, USA. [4]SEARCH (Solutions for Energy, Air, Climate and Health) Center, Yale University, New Haven, CT, USA. [5]Department of Public Health Sciences UConn School of Medicine, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06032-1941, USA. [6]Department of Chemical & Environmental Engineering, School of Engineering and Applied Science, Yale University, New Haven, CT 06511, USA. [7]Analytical Development and Quality Control, Hope Medicine Inc, Shanghai, China. ✉email: anpatton@usfca.edu

Received: 13 June 2022 Revised: 21 October 2022 Accepted: 25 October 2022
Published online: 9 November 2022

International Agency for Research on Cancer (IARC) [3]. Given that $PM_{2.5}$ is produced during combustion, concentrations are often highest in densely-populated urban areas with higher levels of vehicle traffic and fuel combustion at power plants or on more localized scales leading to the potential for variability in $PM_{2.5}$ concentrations over small spatial scales [3].

Within the United States, $PM_{2.5}$ concentrations are required to meet the primary and secondary National Ambient Air Quality Standards (NAAQS) established by the Environmental Protection Agency (EPA) via the Clean Air Act [4]. To ensure that the air quality meets the NAAQS standards, the EPA requires that states operate monitoring sites with high quality sampling equipment that meets a Federal Reference Method (FRM) or Federal Equivalent Method (FEM) in major urban areas. However, there are only 935 $PM_{2.5}$ monitors to cover the entirety of the United States, and of the 25 most populous urbanized areas with a total population of 111 million people, there are only 282 $PM_{2.5}$ monitors [5]. Therefore, the spatial resolution of high quality $PM_{2.5}$ data can be severely lacking for major urban areas. For example, within Baltimore City limits, there is only a single FEM monitor administered by the Maryland Department of the Environment located near the geographic center of Baltimore City during the period of this study [6]. This is highly relevant as intra-city air pollution exposure ranges have been proposed to be as large or larger than exposure ranges between cities and it has been found that there is substantial spatial variation in $PM_{2.5}$ concentrations within a city on a 1–4 km spatial scale [7, 8]. In addition to spatial resolution concerns, gravimetric methods in use at certain FRM stations require 24-hr sampling, sacrificing the ability to measure $PM_{2.5}$ on shorter timescales [6].

In order to fill the spatiotemporal data gaps in air pollution monitoring, low-cost sensor networks have been developed and deployed which consist of many types of sensors that, while less accurate than reference monitors, provide the ability to provide high resolution spatial and temporal measurements relevant at the urban level (e.g. $PM_{2.5}$ concentrations typically below 100 $\mu g/m^3$ with errors on the order of 5 $\mu g/m^3$ as 24-hour averages) [9–11]. One example of a low-cost sensor network is the Solutions for Energy, Air, Climate, and Health (SEARCH) Center's investigation into neighborhood-level variations of air pollutant concentrations in Baltimore, MD [12, 13]. The SEARCH network encompasses low-cost monitors spread across the city measuring PM (mass and number concentrations), ozone, nitric oxide, nitrogen dioxide, carbon monoxide, carbon dioxide, methane, relative humidity, and temperature [12]. However, the reduction in precision compared with an FEM measurement adds complexity to the monitoring such that utilization of the raw sensor data is discouraged without accounting for environmental biases [12, 14]. Therefore, in order to gather sensor data that is both accurate and precise enough for exposure assessment, a combination of field and lab calibration is often recommended to ensure the sensor data is reliable [15, 16]. Lab calibration is both labor intensive and requires laboratory facilities, which is not an option for all low-cost sensor network administrators. However, the presence of an FEM monitor acts as a reference, and co-locating one or more network sensors with the FEM monitor in the field allows for the creation of models that can use raw sensor readings to accurately predict to the reference values. Regardless of a strictly laboratory calibration approach or a combined field and laboratory calibration approach, linear regression is most commonly used to create the calibration model despite known non-linear relationships between $PM_{2.5}$ and meteorological variables [13, 15]. While non-linear models have been developed for sensor calibration, these models still only produce point predictions without estimates of variance [17–20]. However, given that EPA and NIOSH both recommend probabilistic exposure and risk assessments, more accurate assessments are possible if point and/or uniform variance predictions are replaced with predictions from models that also describe variance, particularly on a per-prediction level [21–23].

Therefore, to address the difficulties of laboratory calibration, the lack of fully characterized uncertainty, and known non-linear relationships of predictors, we propose using probabilistic machine learning with gradient boosted decision trees (GBDT) in place of traditional linear approaches for calibration of low-cost $PM_{2.5}$ sensors for sub-city level exposure assessment. While this approach has been conducted prior on indoor occupational exposure based low-cost sensor networks with known schedules and tasks [24], the SEARCH network is outdoors and fully unconstrained by the environmental conditions common in an indoor facility. The GBDT will be trained using raw sensor data directly calibrated to the FEM reference, totally bypassing laboratory calibration, but assuming that the sensors are functional, a check that should be determined prior to deployment. This model will have its accuracy compared to existing linear models which requires laboratory calibrated data. Following model development, we will use the GBDT predictions in a Monte Carlo interpolation approach to create probabilistic $PM_{2.5}$ spatial exposure assessments on the neighborhood scale that can provide health relevant characterization of possible exposures as opposed to a simple deterministic option.

METHODS

Reference data

There is one FEM monitoring site located in Oldtown in central Baltimore and another in Essex on the eastern border of the city. Oldtown lies within the city limits of Baltimore in an area with high traffic density, whereas Essex is outside of the city limits (approximately 15 miles from Oldtown) and is within Baltimore County, the county surrounding Baltimore city. The Oldtown site measures $PM_{2.5}$ on an hourly basis using a Beta Attenuation Mass Monitor, and Essex measures daily average $PM_{2.5}$ once every six days using a gravimetric FRM monitor [6]. Both sites are operated by the Maryland Department of the Environment (MDE) [6].

SEARCH data

The SEARCH data in this study consists of hourly $PM_{2.5}$ measurements from November 2018 through November 2019 taken by 34 separate monitors. Each of the 34 monitors deployed in the network contains a Plantower A003 optical $PM_{2.5}$ sensor as well as a variety of other sensors for gaseous pollutants [25]. Additionally, each monitor has a built-in temperature and relative humidity sensor. Each monitor contains both internal memory storage and a wireless cellular connection via a SIM card that uploads data to a remote server every ten seconds. The locations of the deployed monitors were chosen based on spatial and environmental factors as well as willingness of a property owner to host the monitor. The network has been online since October 2018. The locations of the SEARCH network monitors, Oldtown FEM Monitor (centrally located), and Essex FRM Monitor (eastern coast) are presented in Fig. 1.

There are two SEARCH monitors (B25 and B33) that were co-located with the Oldtown monitor and three monitors (B62, B21, and B8) that were co-located with the Essex monitor (monitor identification numbers are not indicative of the total number of monitors). However, only two of the Essex monitors were ever active at one time. B25 and B33 were deployed from December 2018 through October 2019, and B61, B21, and B8 from February 2019 to August 2019. These monitors will serve as basis for the analysis for the remainder of the calibration analysis.

Modeling

Two separate models were used to model sensor data to reference data. The first was the baseline linear calibration model developed by Datta et al. [13]. This model requires the lab-corrected data to first adjust for non-linear trends, and is shown in Eq. 1.

Equation 1: Linear Regression for $PM_{2.5}$ Calibration to Reference Monitors

$$\begin{aligned} PM_{2.5}Reference = {} & \beta_0 + \beta_1 * RH + \beta_2 * T + \beta_3 * daytime + \beta_4 * weekend \\ & + \beta_5 * RH * PM_{2.5}Sensor + \beta_6 * T * PM_{2.5}Sensor \\ & + \beta_7 * daytime * PM_{2.5}Sensor + \beta_8 * weekend * PM_{2.5}Sensor \end{aligned} \quad (1)$$

The covariates of this model are the lab-corrected low-cost sensor measurement ($PM_{2.5}Sensor$), relative humidity (RH), temperature (T), a

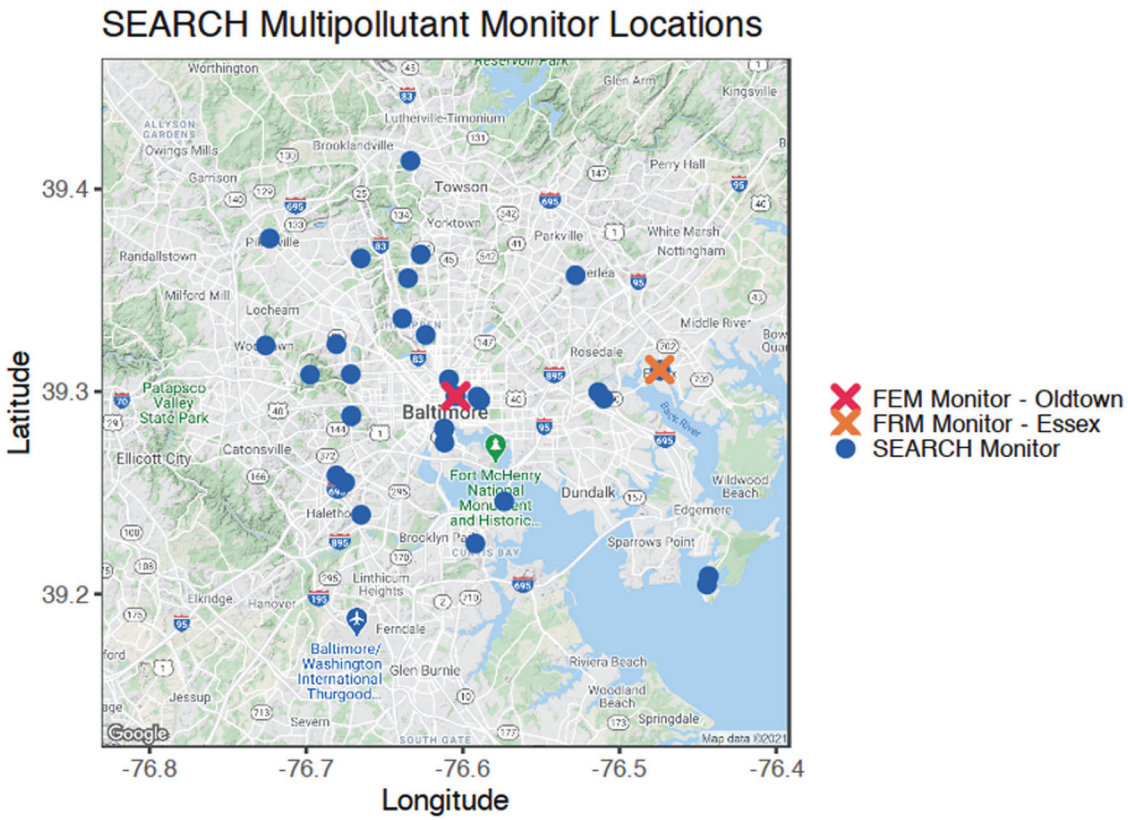

Fig. 1 Map of SEARCH Network Monitors and FEM Monitoring Sites.

binary flag for hours between 7 am and 5 pm (*daytime*), and a binary flag for weekend (*weekend*). $PM_{2.5}$*Reference* refers to the $PM_{2.5}$ as measured by a reference monitor. The model was trained on data where $PM_{2.5}$*Reference* was the measurements from the Oldtown MDE reference monitor, and $PM_{2.5}$*Sensor* was lab-corrected measurements from either monitors B25 or B33. The standard homoscedastic linear regression model is considered which produces constant prediction variances for all predictions. The second model is a gradient boosted decision tree (GBDT) specifically implemented with NGBoost, an open-source probabilistic framework written in python developed by the Stanford Machine Learning Group [26]. Unlike many machine learning methods which are designed for point-prediction of the mean, NGBoost is probabilistic, modeling a distribution for each prediction leading to unique mean and variance. The distribution is specified in the shape of $N(x, \sigma)$. Further specifics on GBDT in general and NGBoost specifically are provided in the Supplement.

Model features. In order to ensure a valid comparison between the two models and demonstrate the efficacy of NGBoost with a small feature space, only five baseline features will be used in each model, four of which (*RH*, *T*, *daytime*, *weekend)* are identical to those from Eq. 1. The only difference between the features of the models is that linear regression uses the lab-corrected $PM_{2.5}$*Sensor* and various interaction terms whereas NGBoost uses the raw, uncorrected $PM_{2.5}$*Raw* sensor data as its fifth feature and does not contain interaction terms. Interaction terms are not specified in the NGBoost model because the structures of decision trees used include them implicitly.

Training and testing datasets

In order to compare and contrast results from the linear models using lab calibrated sensor data presented by Datta et al. [13] with NGBoost using raw sensor data, the time intervals for training and testing from the study will be duplicated with an additional 'monthly' interval added as well. The seven defined training and testing sets are shown in Table 1. While the full, prospective, and three seasonal splits are intended to compare directly to Datta et al. [13] and cover accuracy by season, the Essex split is intended to test the validity of the approach on fully out of sample data and monitors (i.e. not used in the training dataset). However, given that the Essex monitor produces 24-hr average $PM_{2.5}$ concentrations, hourly predictions were made and then averaged up into a 24-hr prediction to allow for comparison against the reference data. This averaging process will compress the values and, as such, the model evaluation metrics for the Essex split should not be directly compared to other train/test splits. The Monthly split is intended to test the model performance on small training set size as well as being to capture some component of sensor drift. Additional specifics on the model training process are available in the Supplement.

Model evaluation

The primary method of evaluation for point (i.e., mean) predictions from the $PM_{2.5}$ models is root mean squared error (RMSE) on the test set. Lower values of RMSE indicate more accurate predictions, and the values have the same units as the predictions and target. The second method of evaluation is the continuous ranked probability score (CRPS) and evaluates the probabilistic results of NGBoost and the confidence intervals of the linear regression. Probabilistic predictions provide more than just a point estimate, and therefore require evaluation of the spread around the point prediction as well as the point prediction itself. For linear regression, each prediction has an identical standard deviation around the mean, whereas NGBoost produces a unique standard deviation for each prediction based on the learned training data. Similar to RMSE, CRPS is in the same units as the predicted variable, with smaller values indicating higher accuracy and takes into account the spread and mean of each prediction distribution [27]. In addition to general evaluation across all predictions, evaluation via RMSE will be performed on bins of reference measurements of 15–20 μg/m^3, 20–25 μg/m^3, 25–30 μg/m^3, and greater than 30 μg/m^3 to evaluate model performance at $PM_{2.5}$ values that are of acute public health concern and often underestimated with linear approaches [13, 28].

Spatial interpolation

In order to use the results of the model to conduct exposure assessments, the calibrated data needs to be spatially interpolated across Baltimore. However, NGBoost's predictions are not simply hourly point predictions at

Table 1. Modeling splits for linear and NGBoost models on oldtown $PM_{2.5}$ data.

Split	Training Set	Testing Set
Full	80% from 12/2018–11/2019	20% from 12/2018–11/2019
Prospective	all from 12/2018–7/2019	all from 8/2019–11/2019
Spring	80% from 3/2019–5/2019	20% from 3/2019–5/2019
Summer	80% from 6/2019–7/2019	20% from 6/2019–7/2019
Winter	80% from 12/2018–2/2019	20% from 12/2018–2/2019
Essex	100% reference data and co-located sensor data from the MDE site of Oldtown 12/2018–11/2019	100% reference data and co-located sensor data from the MDE site of Essex (24 h averages) from 12/2018–11/2019
Monthly	each single month from 1/2019–10/2019	each subsequent month from 2/2019–11/2019

each SEARCH monitor, but a mean and standard deviation of a normal distribution defined as $N_{NGBoost}(x, \sigma)$. Therefore, a resampling process using the distributions as part of the interpolation process was conducted using inverse distance weighting (IDW). IDW operates under the assumption that locations in close proximity are more likely to have similar measurements than those further away, and that the weight of each known measurement in predicting at a location is inversely related to how far away the two are. The general formula for IDW is shown in Eq. 2 with d as distance between the interpolation location and the measured value, i is an unsampled location, z the value at the unsampled location i, and n is the total number of points used in the averaging.

Equation 2: Inverse Distance Weighting (IDW)

$$z_{estimated} = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} z_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \quad (2)$$

The power parameter, p, is used to control the strength of the inverse distance relationship. For larger values of p, more distant measurements are devalued, whereas $p = 0$ corresponds to a straight average across all monitors. While the default selection for p is often 2.0, leave one out cross validation (LOOCV) on the SEARCH monitor mean predictions was conducted to ensure that the power parameter was selected properly and to ensure that interpolation error is propagated through the exposure assessment; additional details are available in the Supplement.

IDW Monte Carlo. Following identification of the optimal power parameter, the next step was to conduct a Monte Carlo simulation using the IDW with the predicted distribution values from the NGBoost model. The Monte Carlo was chosen to run 250 simulations for each hour to balance runtime with accuracy. The steps were as follows:

Step 1: Select a single one-hour portion of the data

Step 2: Select a single draw (i.e., a single value is taken from a distribution) among the 250 draws from each SEARCH monitor's $N_{NGBoost}(x, \sigma)$ prediction from the hour selected

Step 3: Using those single draws, conduct an IDW on a square 256×256 grid (selected for optimal balance of computational speed and resolution) encapsulating Baltimore city limits

Step 4: Following the Monte Carlo, combine all 250 predictions per grid point to obtain the estimated concentration distributions

Following the Monte Carlo simulation, each grid cell's 250 interpolated values were parameterized to a normal distribution following confirmation of normality via a Shapiro-Wilk normality test. Therefore, for each grid location, results were recorded as $N(x, \sigma)_{IDW}$ based on the mean and standard deviation of the interpolated $PM_{2.5}$ values.

Exposure assessment

In order to aggregate an exposure assessment to administrative boundaries, the average of each $N(x, \sigma)_{IDW}$ within the borders of the administrative geometry was defined as the exposure for that zip code, Census Tract, neighborhood, etc. This exposure assessment can also be aggregated temporally, going from single hour bins to days, weeks, or months by averaging the $N(x, \sigma)_{IDW}$ for each hour bin up into the time units desired prior to spatial aggregation.

In order to demonstrate the probabilistic framework, three exposure metrics will be used in an example exposure assessment. The mean and 95th percentile prediction will be provided as more conventional metrics. The third is a threshold-based metric, which represents the probability of exceeding a threshold for a given administrative region and time window. While there are monitors in many of the administrative regions that could theoretically provide single exposure values, the combination of multiple monitors will allow for a complete gradient across the area of interest that can be fit to any scale exposure assessment. Additionally, using multiple monitors in an estimation increases the robustness of the estimate and includes information that takes into account bordering regions. For $PM_{2.5}$, the example threshold was the primary EPA annual standard of 12 μg/m^3 [4]. It is important to note that all three values are produced directly from the probabilistic IDW results of one model. The exposure assessments were conducted on the Community Statistical Area (CSA) level, clusters of similar and known neighborhoods determined by the Baltimore City Planning Department [29].

Software

All modeling was conducted in Python 3.7.7 with spatial analysis and data visualization conducted in R 4.0.2 'Taking Off Again' [30]. The full list of modeling packages, libraries, and their version numbers is provided in Appendix A.

RESULTS

Sensor modeling evaluation

NGBoost outperformed linear regression across all evaluations (Table 2) with an average RMSE of 2.7 μg/m^3 and 3.1 μg/m^3 for NGBoost and linear regression, respectively. NGBoost dramatically outperformed the linear regression in the winter split, with an RMSE of 2.9 μg/m^3 compared to 3.8 μg/m^3, a 30% increase in accuracy in that season. In terms of the probabilistic predictions, NGBoost also had at least a 30% decrease in average CRPS compared to the linear regressions (1.5 μg/m^3 vs. 2.2 μg/m^3), which takes into account both the spread and the mean of the prediction distribution. Therefore, the distribution spread for NGBoost was approximately one third more accurate than the spread for the linear regression model on a per-prediction basis. Crucially, these accuracy improvements were observed even with the fact that the NGBoost was using raw, uncalibrated sensor data as opposed to the lab-corrected data used by the linear regression. Additionally, due to the 24-hr averaging period, the Essex evaluation results had RMSEs that are lower than those of the other evaluation splits. However, the 10% improvement in accuracy of NGBoost compared with linear regression on the fully cross-site out of sample Essex data, demonstrates the transportability of the calibration approach to other monitors.

The predictions from the NGBoost model and linear regression model for the week of February 1, 2019 through February 7, 2019 were compared to the linear regression results and the corresponding MDE reference data in Fig. 2 to provide a visual representation of the efficacy of the non-linear modeling approach. This week in February 2019 was an abnormally poor period for air quality and contained the highest reference measurements in the entire dataset. For comparison, the mean MDE $PM_{2.5}$ during February 2019 was 8.6 μg/m^3, while the daily averages of the first five days of February exceeded 20 μg/m^3. NGBoost generally was able to pick up on the

Table 2. Model evaluation results comparing linear regression with NGBoost across identical training and test splits – RMSE & CRPS.

Split	RMSE (μg/m^3)		CRPS (μg/m^3)	
	Linear Regression	NGBoost	Linear Regression	NGBoost
Full	3.2	2.8	2.3	1.5
Prospective	2.9	2.9	2.2	1.7
Spring	2.6	2.4	1.9	1.4
Summer	2.8	2.6	2.1	1.4
Winter	3.8	2.9	2.6	1.7
Essex (24 h)	2.2	2	-	-

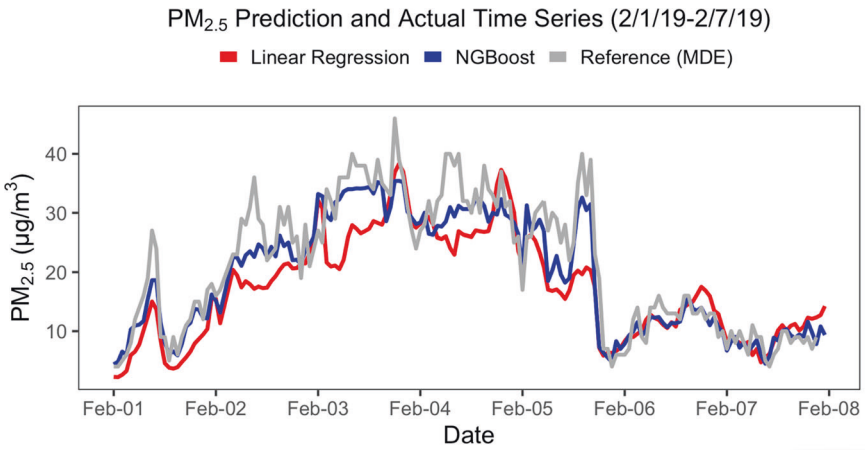

Fig. 2 Time Series for the first week of February, 2019 (highest reference concentration week on record) comparing linear regression (red) and NGBoost (blue) predictions to the MDE reference measurements (gray) where the NGBoost tracks peaks and valleys more effectively than the linear model.

Table 3. Model evaluation results comparing linear regression with NGBoost based on reference measurement range – RMSE.

Reference $PM_{2.5}$ Range (μg/m^3)	RMSE (μg/m^3)			
	Linear Regression	NGBoost	NGBoost Improvement	Count of Hours
15–20	4	3.1	22%	1076
20–25	5.9	4.6	21%	292
25–30	7.4	4.8	35%	114
30+	13.6	9.5	30%	116

peaks and valleys more accurately than linear regression, which is crucially important for health relevant peak exposure events, though the approach still unable to accurately represent all peaks.

Accuracy of peak concentrations. One of the primary positives of the non-linear models is the ability to more accurately represent peak concentrations. In particular with $PM_{2.5}$, hours or days with high measurements are likely to be associated with significant negative public health impacts. NGBoost demonstrated significantly reduced error during high exposure hours compared to linear regression (based on full model training/test split). The RMSE for four categories of reference exposure are shown in Table 3 along with the corresponding improvement of NGBoost compared to the traditional linear approach based on the predictions of the fitted model across the entire dataset. Furthermore, the average residual for linear regression at reference concentrations greater than 15 μg/m^3 was −4.33 μg/m^3 whereas for NGBoost it was −2.12 μg/m^3, demonstrating an approximately 50% reduction in the negative bias at high reference concentrations.

Exposure assessment

IDW predictions were created for every hour from February 2019 through November 2019 using the identified optimal power parameter of $p = 2.0$. An example of a single day exposure assessment on the CSA level within Baltimore city limits was conducted on June 5, 2019 to highlight the spatial variability observed using the network. June 5, 2019 was selected as it had the largest difference in single-day mean CSA predictions (approximately 10.3 μg/m^3) between any two monitors, considering days with more than 20 monitors in operation. This gradient between monitors is not captured with a single reference instrument like the Oldtown FEM monitor. Mean and 95th percentile $PM_{2.5}$ values are shown for June 5, 2019 in Fig. 3a. For comparison, August 1, 2019 (Fig. 3b) had a maximum CSA predicted difference of 3.5 μg/m^3. In both cases, although more pronounced in on June 5, there is a region of high concentrations in the center of the city, with additional high concentration areas in the northeast and northwest areas, likely corresponding to commuting traffic on major interstates I-83, which runs north-south through the center of the city and I-695

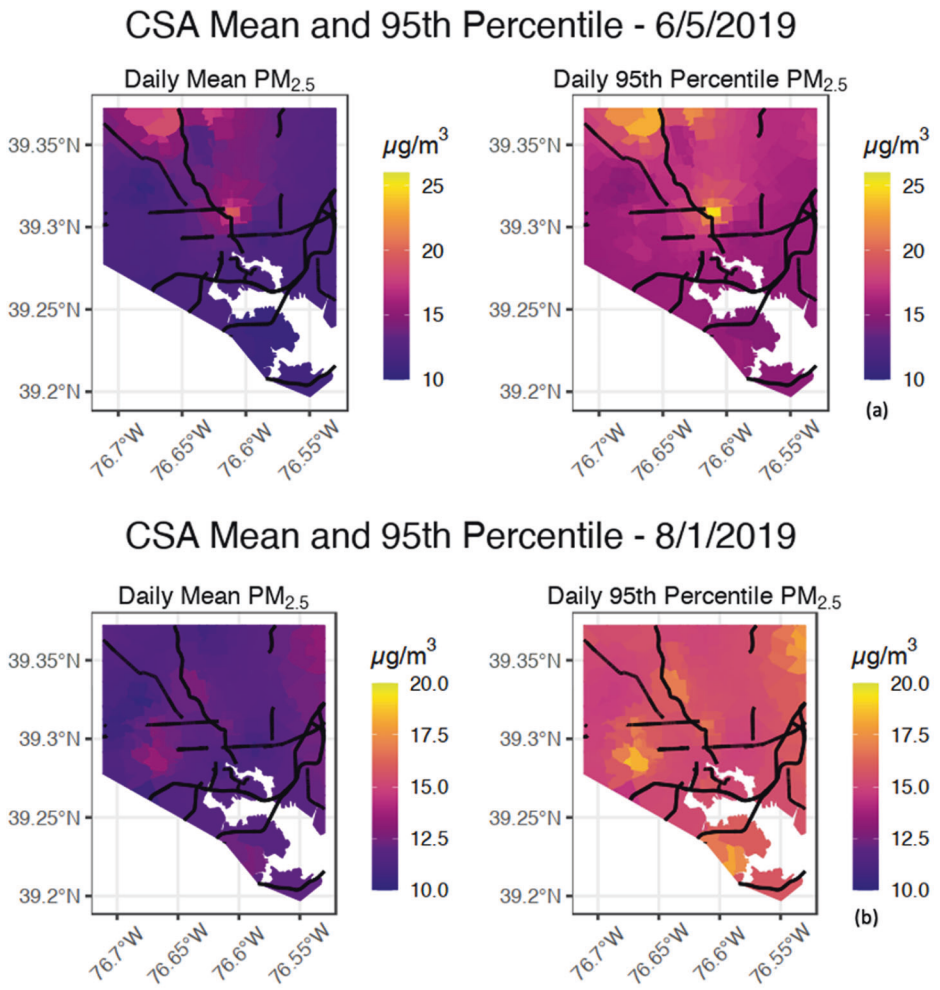

Fig. 3 Mean (left) and 95th Percentile (right) $PM_{2.5}$ exposures by Community Statistical Association (CSA) for June 5, 2019 (**a**) and August 1, 2019 (**b**) with major roads and highways denoted.

beltway that surrounds the city slightly outside the city limits (not shown).

Therefore, the probabilistic nature of the NGBoost outputs, predicting the mean and standard deviation for each prediction enables a mapped representation of the probability of any location exceeding a specified threshold (e.g. 12 μg/m^3). An example of such a map for the probability that any location exceeds 12 μg/m^3 for the 24-hr period on June 5, 2019 as shown in Fig. 4. This is a powerful approach to assess spatially-resolved risk for exceeding threshold values in a complex urban landscape using low-cost distributed measurement networks. Of the 278 CSAs in Baltimore on June 5, 2019, 158 had a greater than 50% chance of exceeding 12 μg/m^3 and 28 had a greater than 90% chance of exceeding 12 μg/m^3. The CSAs with the highest exceedance probabilities are the center city areas near major commuting intersections (76.6 °W, 39.3 °N), and the northern areas that are adjacent to major interstate traffic (north and northwest borders), however this is only one possible explanation for June 5th exposures, as traffic is not the only source or causative factor for $PM_{2.5}$ exceedances. Additionally, the flexibility of the approach can be seen in Fig. 5 which has been aggregated over 316 days, uses Census Tracts as the geographic aggregation unit, and a concentration of interest of 10 μg/m^3.

In comparison to Fig. 4, there are differing patters on spatial variability in exposure in Fig. 5, highlighting the point that the small-scale variation in the daily or sub-daily scale maps are not simply just duplicates of the annual or long-term exposure pattern. Of note, in both Fig. 4 and Fig. 5, large city parks in the northwest (76.7 W, 39.3 N) and center east (76.6 W, 39.3 N) of the city show up as low exposure zones compared to their surrounding developed areas.

DISCUSSION

The use of machine learning for predictive purposes in air pollution sensor data has seen substantial growth in the last several years. Large-scale approaches often utilize satellite data, country scale sensor networks, land use data, topography, etc. and have been built using random forests, GBDTs, and neural nets [31–34]. On a smaller scale more analogous to SEARCH, personal monitoring device networks, mobile sampling networks, and city-scale sensor networks have also demonstrated the utility of machine learning regression techniques to optimize predictions and take into account environmental factors [17–20]. However, while prediction of $PM_{2.5}$ using sensor measurements and additional data has been conducted by numerous studies, this study fills a unique position by providing a methodology for both increasing the utility of low-cost sensor networks by creating a probabilistic output useful for exposure assessments, a state-of-the-art model that improves on existing approaches, and also removes the need for lab-calibrated data, a time intensive process for mitigating environmental biases for $PM_{2.5}$ data.

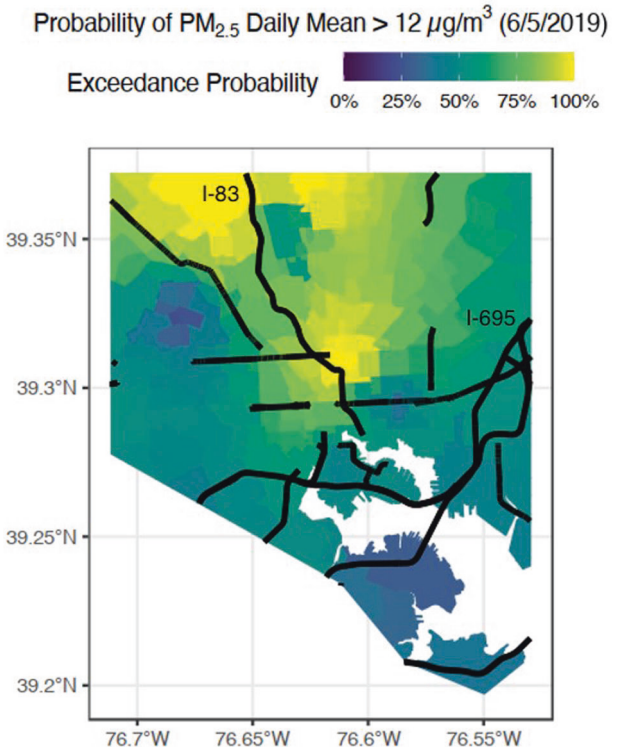

Fig. 4 Probability of daily mean $PM_{2.5}$ exceeding 12 mg/m^3 by CSA on June 5, 2019 with major highways and roads.

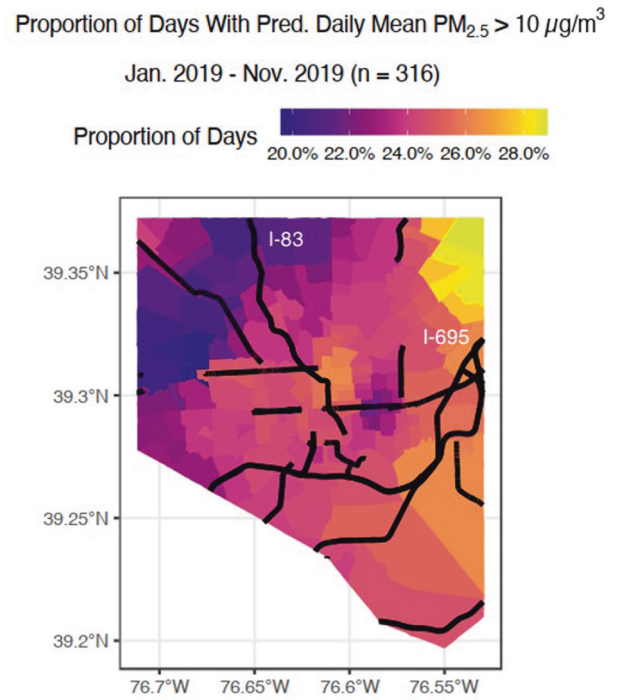

Fig. 5 Proportion of days with a predicted daily mean $PM_{2.5}$ exceeding 10 mg/m^3 by Census Tract from January 2019 – November 2019 with major highways and roads.

The Plantower PMS A003 sensors used in this work produce $PM_{2.5}$ readings by scattering laser light on drawn in air which is used to count particle sizes which are then converted into concentrations. This raw sensor can be lab-calibrated (accounting for known environmental biases), a time and resource intensive process, in order to ensure accuracy and precision [15, 16]. The SEARCH network is also deployed in the region of a Maryland Department of the Environment $PM_{2.5}$ monitoring station which measures $PM_{2.5}$ using a reference Federally Equivalent Method (FEM) $PM_{2.5}$ measurement [5, 6]. In previous studies, co-located SEARCH sensor monitors with the FEM monitor were used to develop a linear regression model that used lab-calibrated sensor data, temperature, relative humidity, weekend (binary), and daytime (binary) to model gold standard $PM_{2.5}$ [13]. Although temperature and relative humidity are known parameters of concern when measuring $PM_{2.5}$, they have established non-linear relationships with the ultimate $PM_{2.5}$ measurement as well as each other, and these non-linear relationships and large sensor to sensor variability are the reasons why lab calibration is often necessary [13, 15]. Alternately, in order to capture the non-linear relationships without lab calibration, gradient boosted decision trees (GBDT), a popular tool for non-linear regression were used and showed that they were more accurate than linear regression without the calibration step. In addition, the NGBoost specific GBDT utilized in this study was probabilistic, producing unique means and standard deviations for each prediction output, in contrast to the linear regression which provided uniform standard deviations for calibration uncertainties across all predictions. The unique means and standard deviations from NGBoost resulted in a nearly 30% decrease in CRPS compared to a uniform variance uncertainty from linear regression.

While creating models that produce accurate probabilistic predictions is interesting from an academic perspective, it is the application of the models that can result in actionable data products, as seen in Fig. 4, that presents exceedance probabilities for relevant regulatory/health protective standards and facilitates comparisons between neighborhoods for environmental justice applications. The flexibility of the approach is also crucially important for exposure assessments, as the data can be aggregated to whichever administrative boundary the researcher prefers for their problem, as well as whatever timescale is required. Exposures can also be left as a continuous gradient and used for nearest point analysis for studies requiring fine scale exposures linked to residences, places of work, etc. Leveraging the probabilistic prediction optimized NGBoost allows for the use of the distribution for further analysis such as probabilistic risk assessments, more accurate best and worst case scenarios, and any other situation where a parameterized distribution would be more useful than a point prediction, particularly since Patton et al. [24] showed that the use of probabilistic exposure estimates more accurately estimate upper bound exposures from low-cost sensor networks than the use of a point estimate. Approximations such as using the 95th percentile prediction from NGBoost would approach a worst-case scenario, but one that is modeled with a unique mean and standard deviation based on the input data. This is in contrast to a linear regression where a 95th percentile is based on a uniform standard deviation across all data points. Therefore, our approach allows for low probability occurrences such as the 95th percentile outcome to be modeled with precision.

In terms of limitations, the specific tuned and fitted models covered in this paper are not universally applicable. The intent was to provide a framework for other investigators to use this approach on their own sensor networks and pollutants. Unique models should be tuned and fitted for every application, which is both a potential limitation but also lends itself to highly customizable solutions. Furthermore, the features for NGBoost in this setting were not engineering or optimized but were simply the same as what was identified to be optimal for the linear regression by Datta et al. [13] to facilitate comparability. Therefore, it is likely that specific feature engineering would yield increased model accuracy. Finally, we have a small validation set comprised

only of the state's compliance monitoring to evaluate our results, with no validation at other locations.

Further research into these methods should consider the addition of more reference-sensor pairs that would allow for features that more completely characterize the local environment of each pair. For example, adding land use, topographic, or traffic features would easily be possible with our approach. While adding regulatory monitoring sites is not feasible, a short-term high cost/accuracy instrument could be co-located with several monitors to provide reference data across the entire network. In addition to the potential for an expanded feature space, one of the primary adjustments to make is to determine the amount of training data needed. While this will vary by pollutant, features, model choice, and prediction quality desired, capturing climate variation across several months would be recommended. Further, it is possible that an ensemble of high bias low variance linear models (not likely to overfit, but likely overly generalized) and low bias high variance GBDT (possibility of overfit, but not overly generalized) would be useful in a setting where a limited amount of training data was available with no option to acquire more. Lastly, it is possible that a temporal weighting feature that weights newer data more heavily would additionally yield increased accuracy as a means to combat sensor drift—methods such as error optimized exponential weighting would be an option [15].

The framework for converting uncalibrated $PM_{2.5}$ sensor data into a probabilistic exposure assessment using probabilistic gradient boosted decision trees captures the non-linearity of the relationship between $PM_{2.5}$, relative humidity, and temperature, while providing more accurate and more useful probabilistic and deterministic output. The exposure assessments derived from the probabilistic modeling allows for small scale understanding of $PM_{2.5}$ exposure and variability that can be of use in acute and sub-chronic epidemiological studies.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are not publicly available due to them being a part of continuous ongoing research but are available from the corresponding author on reasonable request.

REFERENCES

1. World Health Organization. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. 2018.
2. Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet. 2017;389:1907–18.
3. International Agency for Research on Cancer. Outdoor Air Pollution (Vol. 109). Lyon; 2016.
4. Environmental Protection Agency. NAAQS Table. 2010. Available from: https://www.epa.gov/criteria-air-pollutants/naaqs-table
5. Apte J, Messier K, Gani S, Brauer M, Kirchstetter T, Lunden M, et al. High-resolution air pollution mapping with Google Street View cars: exploiting big data (Supplemental Material). Environ Sci Technol. 2017;51:6999–7008.
6. Maryland Department of the Environment. Ambient Air Monitoring Network Plan for Calendar Year 2019. Baltimore; 2018.
7. Ye Q, Li HZ, Gu P, Robinson ES, Apte JS, Sullivan RC, et al. Moving beyond fine particle mass: High-spatial resolution exposure to source-resolved atmospheric particle number and chemical mixing state. Environ Health Perspect. 2020;128.
8. Saha PK, Sengupta S, Adams P, Robinson AL, Presto AA. Spatial correlation of ultrafine particle number and fine particle mass at urban scales: implications for health assessment. Environ Sci Technol. 2020;54:9295–304.
9. Snyder EG, Watkins TH, Solomon PA, Thoma ED, Williams RW, Hagler GSW, et al. The changing paradigm of air pollution monitoring. Environ Sci Technol. 2013;47:11369–77.
10. Piedrahita R, Xiang Y, Masson N, Ortega J, Collier A, Jiang Y, et al. The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. Atmos Meas Tech. 2014;7:3325–36.
11. Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics. 2009;21:n/a–n/a.
12. Buehler C, Xiong F, Levy Zamora M, Skog K, Kohrman-Glaser J, Colton S, et al. Stationary and portable multipollutant monitors for high spatiotemporal resolution air quality studies including online calibration. Atmos Measurement Tech. 2020;in review.
13. Datta A, Saha A, Zamora ML, Buehler C, Hao L, Xiong F, et al. Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmos Environ. 2020;242:117761.
14. Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, et al. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Vol. 116, Environment International. Elsevier Ltd; 2018. 286–99.
15. Levy Zamora M, Xiong F, Gentner D, Kerkez B, Kohrman-Glaser J, Koehler K. Field and laboratory evaluations of the low-cost plantower particulate matter sensor. Environ Sci Technol. 2019;53:838–49.
16. Borrego C, Ginja J, Coutinho M, Ribeiro C, Karatzas K, Sioumis T, et al. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise – Part II. Atmos Environ. 2018;193:127–42.
17. Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos Environ. 2017;151:1–11.
18. Loh BG, Choi GH. Calibration of portable particulate matter–monitoring device using web query and machine learning. Saf Health Work 2019;10:452–60.
19. Lim CC, Kim H, Vilcassim MJR, Thurston GD, Gordon T, Chen LC, et al. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ Int. 2019;131:105022.
20. Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, et al. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech. 2018;11:291–313.
21. EPA. Risk Assessment Forum White Paper: Probabilistic Risk Assessment Methods and Case Studies. 2014. Available from: https://www.epa.gov/sites/production/files/2014-12/documents/raf-pra-white-paper-final.pdf
22. NIOSH. How NIOSH Conducts Risk Assessments. 2017. Available from: https://www.cdc.gov/niosh/topics/riskassessment/how.html
23. Daniels R, Gilbert S, Kuppusamy S, Kuempel E, Park R, Pandalai S, et al. Current Intelligence Bulletin 69 - NIOSH Practices in Occupational Risk Assessment. 2020.
24. Patton AN, Medvedovsky K, Zuidema C, Peters TM, Koehler K. Probabilistic machine learning with low-cost sensor networks for occupational exposure assessment and industrial hygiene decision making. Ann Work Exposures Health. 2022;66:580–90.
25. Buehler C, Xiong F, Zamora ML, Skog KM, Kohrman-Glaser J, Colton S, et al. Stationary and portable multipollutant monitors for high-spatiotemporal-resolution air quality studies including online calibration. Atmos Meas Tech. 2021;14:995–1013.
26. Duan T, Avati A, Ding DY, Thai KK, Basu S, Ng AY, et al. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. 2019.
27. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102:359–78.
28. Heffernan C, Peng R, Gentner DR, Koehler K, Datta A. Gaussian Process filtering for calibration of low-cost air-pollution sensor network data. arXiv. 2022 [cited 2022 Jun 7]. Report No.: arXiv:2203.14775. Available from: http://arxiv.org/abs/2203.14775
29. Baltimore City Department of Health. Neighborhood Health Profiles - Frequently Asked Questions | Baltimore City Health Department. 2017 [cited 2020 Sep 30]. Available from: https://health.baltimorecity.gov/node/231
30. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
31. Chang FJ, Chang LC, Kang CC, Wang YS, Huang A. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. Sci Total Environ. 2020;736:139656.
32. Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, et al. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environ Pollut. 2018;242:675–83.
33. Zhan Y, Luo Y, Deng X, Grieneisen ML, Zhang M, Di B. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ Pollut. 2018;233:464–73.
34. Zhao Z, Qin J, He Z, Li H, Yang Y, Zhang R. Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. Environ Sci Pollut Res. 2020;1–18.

ACKNOWLEDGEMENTS

This manuscript has not been formally reviewed by the Environmental Protection Agency (EPA). The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. The EPA does not endorse any products or commercial services mentioned in this publication. Further, any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

AUTHOR CONTRIBUTIONS

AP: conceptualization, methodology, software, formal analysis, data curation, writing—original draft, writing—review & editing, visualization. AD: Methodology, software, formal analysis, data curation, writing—original draft, writing—review & editing. MLZ: Data curation, writing—review & editing, investigation. CB: writing—review & editing, investigation. FX: writing—review & editing, investigation. DG: writing—review & editing, investigation, funding acquisition. KK: conceptualization, methodology, data curation, writing—original draft, writing—review & editing, supervision, funding acquisition.

FUNDING

This publication was developed under Assistance Agreement no. RD835871 awarded by the U.S. Environmental Protection Agency to Yale University. AP was supported by a grant from the U.S. Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health to the Johns Hopkins Education and Research Center for Occupational Safety and Health (award number T42 OH0008428). AD was supported by the National Science Foundation DMS-1915803 and the National Institute of Environmental Health Sciences (NIEHS) grant R01ES033739. CB was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1752134. DG and FX would also like to acknowledge support from HKF Technology and Ken Hu. MLZ was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under award numbers K99ES029116 and R00ES029116.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41370-022-00493-y.

Correspondence and requests for materials should be addressed to Andrew Patton.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.