Electronic Supplementary Material (ESI) for Environmental Science: Atmospheres.

Supporting Information:

Using Spatio-temporal Prediction Models to Quantify $PM_{2.5}$ Exposure due to Daily Movement

Authors: Sakshi Jain[†], Albert Presto[‡], Naomi Zimmerman[*,†]

†Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada

‡Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, USA

2 Tables, 7 Figures, 14 pages.

*Corresponding author:
Professor Naomi Zimmerman
Dept. of Mechanical Engineering
The University of British Columbia
2054-6250 Applied Science Lane
Email: nzimmerman@mech.ubc.ca
Tel. 604-822-9433
Fax. 604-822-2403

S1. Data Collection

For this work, a previously published prediction model data from a network of 47 RAMPs deployed in Allegheny County (Pennsylvania, USA)[4] was used to estimate concentration at every 50 x 50 m grid in the City of Pittsburgh. The locations of RAMPs and EPA monitors are shown in Figure S1. RAMPs contain either a Met-One Neighborhood Monitor (Met-One) or a PurpleAir PA-II (PPA). In this network, the default sensor was the Met-One (40/47 sites); PurpleAir monitors were only used in cases where a Met-One monitor was offline or otherwise unavailable.

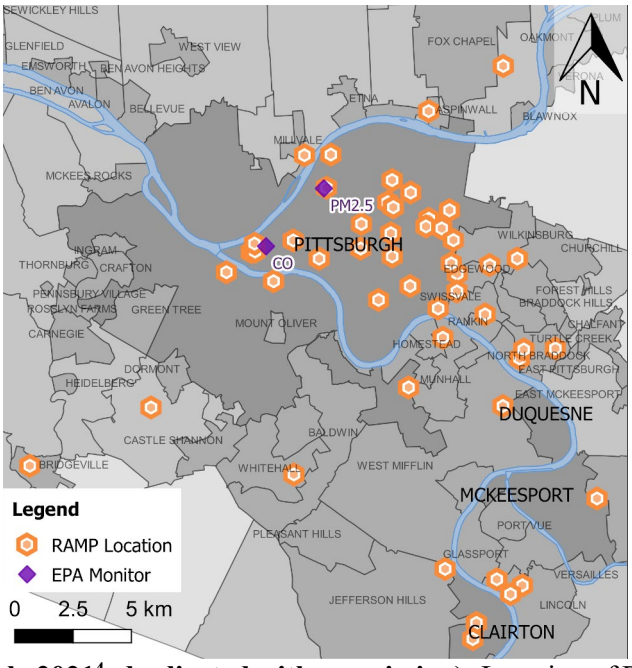

Figure S1 (from Jain et. al., 2021[4], duplicated with permission): Location of RAMPs (orange hexagons) and EPA monitors (purple diamonds) across Allegheny County.

On average, across 47 sites, 180 days of data was used; the period of observation (filtered and calibrated data) across each $PM_{2.5}$ sensor is displayed in Figure S2 (green boxes). Calibrated 15-minute data can be found at https://doi.org/10.5281/zenodo.8264657.

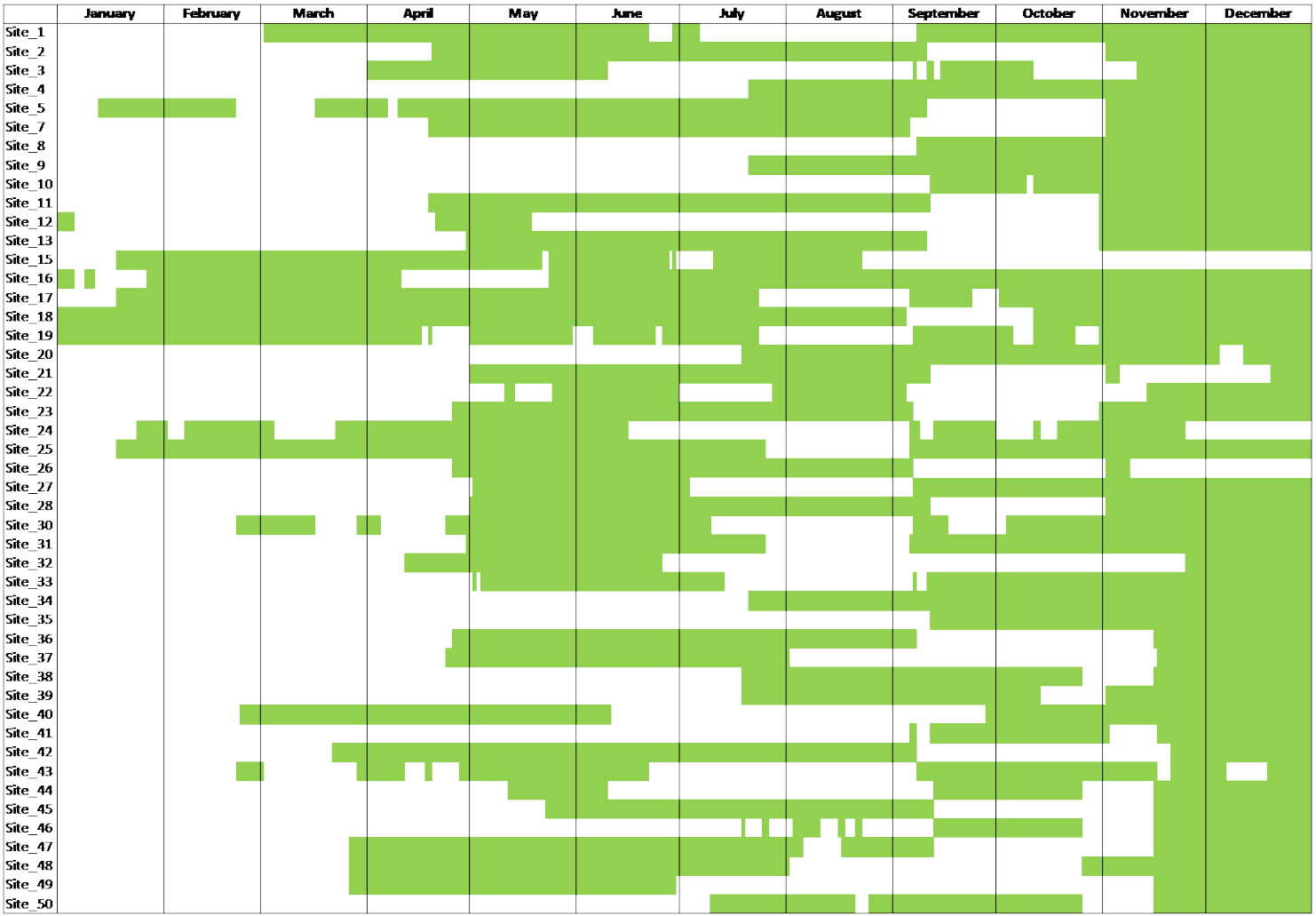

Figure S2: Filtered and calibrated data used for this work (green boxes) across each site over the duration of study period (January-December 2017). The number of days of data range from 95 (site 12) to 328 (site 18), with an average of 180 days of data.

S2. QA/QC for RAMP data

a) Data Cleaning

RAMP data underwent both automated and manual cleaning. The automated cleaning involved two major bounds: (1) Limits for $PM_{2.5}$ collected was set as 0-1000 $\mu g/m^3$, and readings with concentrations outside the limits were removed and (2) a change faster than 50 $\mu g/m^3$ per minute was filtered as it was deemed unphysical. The manual cleaning of RAMP data included a review of RAMP data and post-deployment checks of the instruments, and removing or re-scaling the data-point (e.g., flat-lining).

b) Calibration

The calibration of data used in this work was performed using methodologies developed by Malings et al. (2020)[1]. A subset of methodology is paraphrased below. All methodologies, results, figures and tables in this section are the intellectual property of Carl Malings and have been paraphrased or reproduced here with his permission.

RAMPs contain either a Met-One Neighborhood Monitor (Met-One) or a PurpleAir PA-II (PPA) that were calibrated against Beta Attenuation Monitors (BAM, a Federal Equivalent Method) at Allegheny County Health Department (ACHD) or Pennsylvania Department of Environmental Protection (DEP). Met-One sensors were calibrated using the seasonally-varying hygroscopic growth correction formula (linear correction) and PPA sensors were calibrated using the empirical approach (Malings et. al., 2020[1]).

Met-One Sensor calibration:

The following equation was adopted to calibrate Met-One sensors to the reference grade instrument.

$$[\text{corrected } PM_{2.5}] = \theta_1 \frac{[PM_{2.5} \text{ as reported}]}{fRH(T, RH)} + \theta_0$$

Here, corrected $PM_{2.5}$ is the reference concentration, and $PM_{2.5}$ as reported is the RAMP concentration. Values of coefficients (θ_0 and θ_1) are reported in Table 1.

Table S1: Coefficients calculated using typical linear regression techniques for Met-One sensors, reproduced from Malings et. al., 2020[1].

	θ_0		θ_1	
	Coefficient	S.D.	Coefficient	S.D.
Summer	5.28	0.09	1.5	0.01
Winter	2.03	0.08	1.5	0.01
Other	1.68	0.13	1.76	0.02

PPA sensor calibration:

$$[\text{corrected } PM_{2.5}] = \begin{cases} \beta_0 + \beta_1(PM_{2.5}) + \beta_2 T + \beta_3 RH + \beta_4 DP(T, RH) \text{ if } [PM_{2.5}] > 20\ \mu g/m^3 \\ \gamma_0 + \gamma_1(PM_{2.5}) + \gamma_2 T + \gamma_3 RH + \gamma_4 DP(T, RH) \text{ if } [PM_{2.5}] \leq 20\ \mu g/m^3 \end{cases}$$

Here, corrected $PM_{2.5}$ is the reference concentration, and $PM_{2.5}$ is the RAMP concentration. Coefficients (β and γ) can be found in Table S4 of Malings et al. (2020)[1].

S3. Limit of Detection (LOD)

For this work, LOD is the smallest concentration that can be reliably measured by the low-cost sensors and was identified as 5-µg/m^3. We opted against removing the datapoint below LOD as the data would then be skewed high. Similarly, replacing the datapoints below LOD with 0 would result in data skewed low. Therefore, we opted for replacing the data measured below LOD with 3.53-µg/m^3 (LOD/√2), as recommended by Hornung and Reed (1990)[2] and Tekindal (2015)[3].

Figure S1 shows a boxplot for the percent of data at each of the 47 sites replaced due to being below LOD.

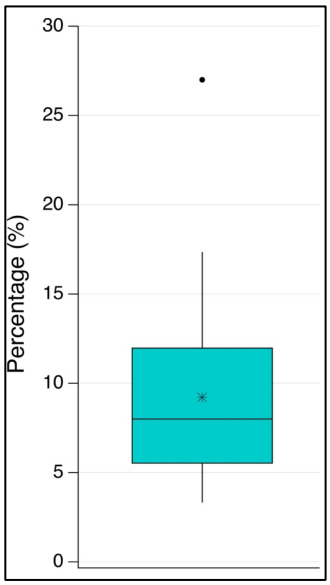

Figure S3: Boxplot for percent of data at each site (n = 47 sites) that were below Limit of Detection (LOD; 5-µg/m^3) and replaced with LOD/√2 (=3.53-µg/m^3).

S4. Selection of prediction models and variables

In our previous study (Jain et al., 2021[4]), we compared regression, random forests, and hybrid regression-random forest models for both standard and time-decomposed signals. In time-decomposed analysis, we created separate land use regression / random forest models for persistent enhancements (> 8 h duration), longer-lived events (2-8 h duration) and short-lived events (<2 h duration) and layered these on top of the regional background to predict overall $PM_{2.5}$ concentrations.

For modeling of $PM_{2.5}$, random forest models outperformed regression models (R^2 increases of 0.17-0.19; normalized mean absolute error decreased by 4-7%). Hybrid regression-random forest models were created to address the incapability of random forests to extrapolate. However, we found that hybrid models didn't improve the overall model performance or robustness of random forest models. Therefore, random forest models were chosen for predicting concentrations at every grid cell.

For random forest models, the standard signal model had comparable performance to decomposed signals. However, decomposing the signal improved the relative importance of static (spatial) variables in the model for short-lived events. This implies that local spikes in concentrations can be predicted using land use characteristics in the nearby locations. Since we are primarily looking into spatial effects for this work, we opted for the decomposed signal model. For a detailed discussion of the relative model performance of land use regression vs. land use random forest, as well as the impact of time decomposition, see Jain et al., 2021[4].

Table S2: Top 5 most important variables for modeling of random forest decomposed signal. Value in the brackets signify buffer distance.

Signal	**Spatial variables**	**Temporal variables**
Persistent enhancement	Population density (100m), road length (100m), housing density (100m), rail length (100m)	EPA's daily $PM_{2.5}$ measurements
Long-lived events	Elevation, vehicle density (50m), bus fuel consumption (50m), inverse distance to the road	EPA's daily CO measurements
Short-lived events	Elevation, road length (50m), vehicle density (100m), bus fuel consumption (100m)	Wind

S5: Modeling

The methodology adopted for building regression models was performed by Jain et al. (2021)[4] and paraphrased below in steps.

1) Concentrations below LOD (5-μg/m^3) were replaced with LOD/√2.
2) Wavelet decomposition was performed on calibrated 15-minute calibrated data and resulted in three signal components – short-lived events (<2 hours), long-lived events (2-8 hours) and persistent enhancements (> 8 hours). Regional background concentrations were dynamically calculated as the lowest persistent enhancement across 47 locations at any given time.
3) The four signals were each averaged into daily average concentrations.
4) LURF predictions were built for each component of the signal (short-lived, long-lived and persistent enhancements) using various spatial and temporal variables and tested for validation using the leave-one location-out cross-validation (LOLOCV). The spatial and temporal candidate variables can be found in Table S3 of Jain et al. (2021)[4] and the top five most important variables can be found in Table S2.
    a. For LURF model building on training dataset, 10-fold cross-validation was used, with termination at 1000 trees and minimum terminal node size of 1.
    b. Initial model was created using all the spatial and temporal variables, and variable importance factor (VIF) for each predictor variable was reported.
    c. Multiple models were built, each one by removing the least important variable from the model and the new R^2 was reported. For each iteration, VIF from the original model was used.
5) The decomposed signal predictions were calculated by adding the predictions: $LURF_{short\text{-}lived}$ + $LURF_{long\text{-}lived}$ + $LURF_{persistent}$ + regional background.

S6. Land Use Types by Allegheny County GIS Group[5]

1. Water
2. Transportation
3. Forest
4. Grasslands
5. Agriculture
6. Low-density residential
7. Medium-density residential
8. High-density residential
9. Identified malls
10. Commercial
11. Light industrial
12. Heavy industrial
13. Strip mine
14. Non-vegetative

S7. Spatial distribution at 100m buffers for residential and commercial areas

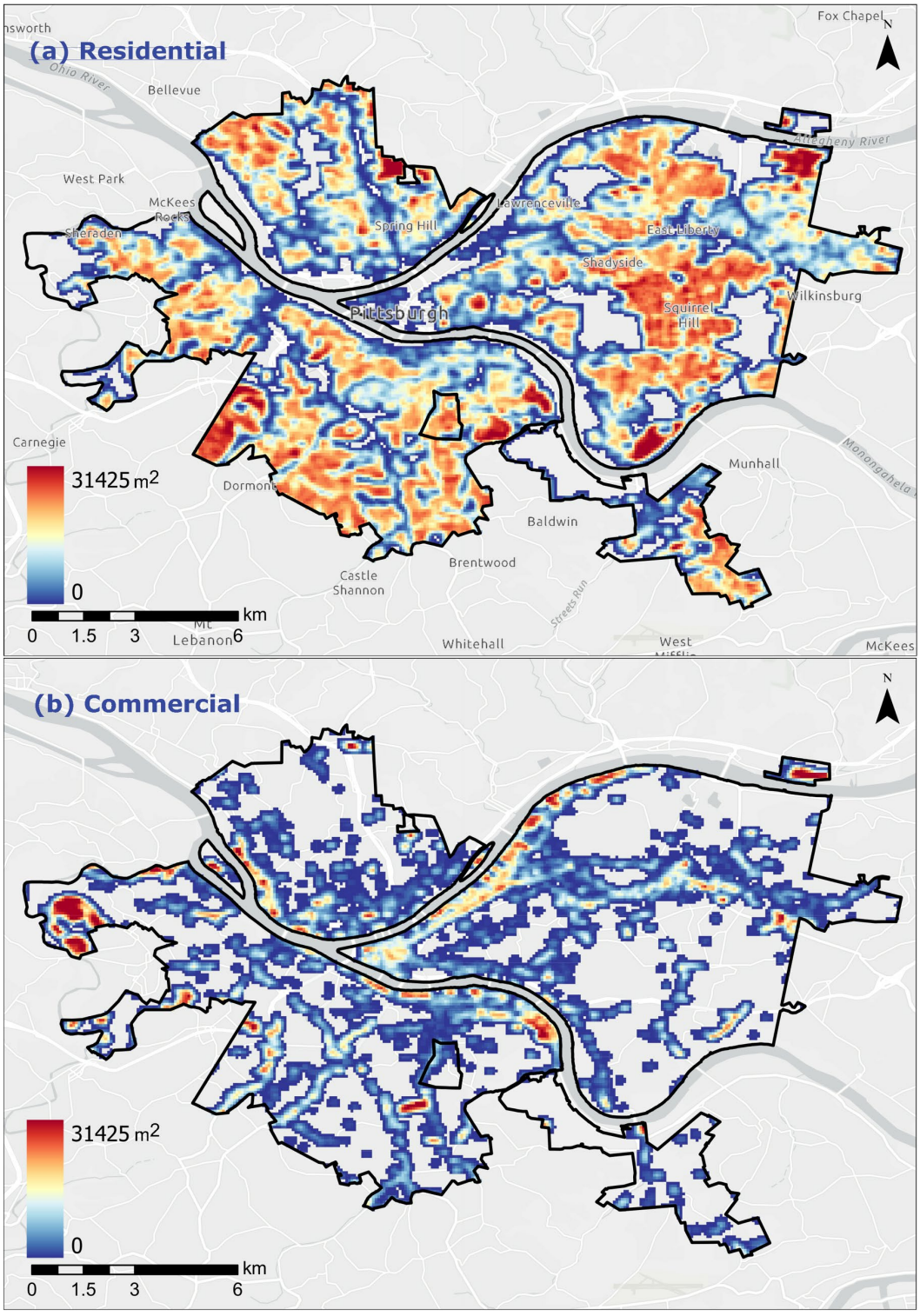

Figure S4: Spatial distribution of (a) Residential and (b) Commercial areas at 100m buffers in Pittsburgh city, obtained via Allegheny County GIS Group[5]. Grid cells with no value (colorless) imply residential or commercial density is zero for that grid.

S8. Average day-wise concentrations in 2017

This work segregates the daily concentrations into weekday (Monday through Friday) and weekend (Saturday and Sunday) concentrations. Figure S3 shows a boxplot for EPA's daily $PM_{2.5}$ concentrations across different days of the week for 2017. Amongst weekdays, Mondays through Thursdays have similar medians (~8 µg/m^3). Median Friday concentrations are higher (~9 µg/m^3) and can be attributed to higher rate of various evening activities by individuals or citywide events (e.g., dining out, game nights, festivals). Nonetheless, we opted to group weekday concentrations because we expect similar behavioural movement (e.g., amount of time spent) between residential and commercial areas during the weekdays. Analogously, even though Saturday and Sunday concentrations are also dissimilar, we have grouped them together into weekends.

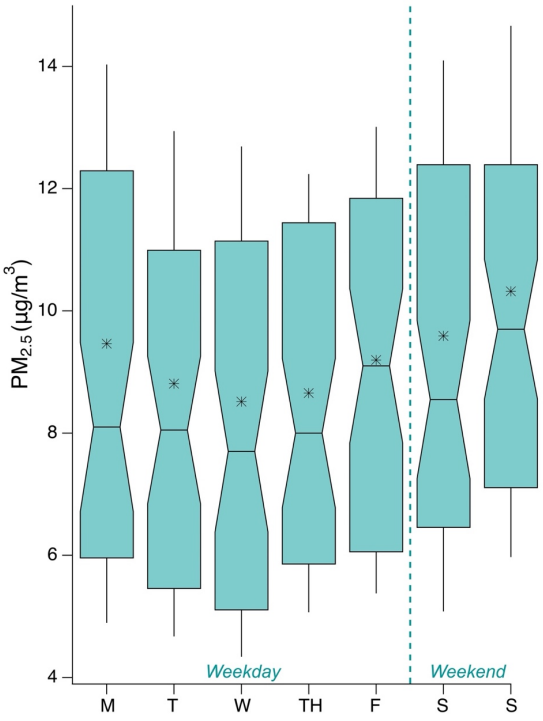

Figure S5: Boxplot for mean day-wise concentrations in 2017 for data collected at EPA's Lawrenceville site in Pittsburgh city.

S9. Uncertainties in measurement and models

We identified two uncertainties associated with this work – uncertainties in measurements taken by RAMPs and uncertainties in prediction modeling.

- Uncertainties in measurements: In Pittsburgh city, we collocated a RAMP at EPA's Lawrenceville site. We compared calibrated daily $PM_{2.5}$ measurements from RAMP with EPA's data and found normalized mean error to be 18%.

$$Normalized\ mean\ error\ (ME) = \frac{\sum_{i=1}^{n}(EPA\ measurement_i - RAMP\ measurement_i)/n}{Average\ EPA\ measurements}$$

- Uncertainties in modeling: Random Forest models are created as an ensemble of decision trees. However, these decision trees use the mean value estimate predicted values. To ascertain that we account for uncertainties associated at this stage of modeling, in addition to mean values, we noted 2.5th and 97.5th percentile predicted values. Using these, we found normalized mean error to be about 28% and 72% respectively for 2.5th and 97.5th percentile predicted values.

We assume that uncertainties in modeling had a higher overall effect due to higher normalized mean error. Therefore, we opted to evaluate model uncertainties, and used the predicted values across 5th and 95th percentile random forest models and compared the concentrations in residential and commercial areas (Figure S6).

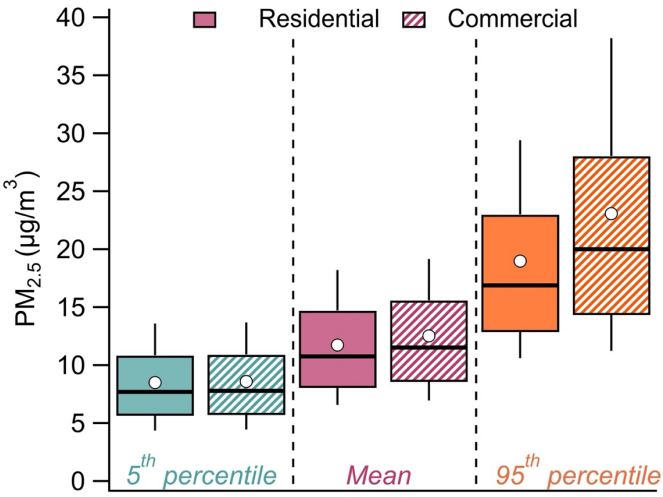

Figure S6: Boxplots for daily predicted $PM_{2.5}$ for residential (plots with solid colors) and commercial (plots with diagonal lines) land-use type separately. Blue and orange boxplots refer to annual average predicted $PM_{2.5}$ concentrations when random forest models noted 5th and 95th percentile concentrations, instead of mean concentrations (pink boxplots).

Average and median concentrations in commercial areas were higher than concentrations at residential areas across the 5th percentile, mean and 95th percentile of the random forest models (Figure S6). For the mean (pink boxes; Figure S6), both median and average concentration at commercial areas were 6% higher than at residential areas. For the 5th percentiles of the random forest models (blue boxes, Figure S6) both average and median concentration at commercial areas were 1% higher than residential areas. Similarly, for the 95th percentiles of the random forest models (orange boxes, Figure S6), average and median concentration at commercial grids were 22% and 18% higher. Although the absolute value between models (5th, mean and 95th percentile; Figure S6) are different, the average concentration in commercial areas is always higher when compared to residential areas. As such, addressing these uncertainties strengthens our argument that average $PM_{2.5}$ concentrations that the population experiences may be underreported when only residential address is considered.

Section S10: Static and Dynamic Models

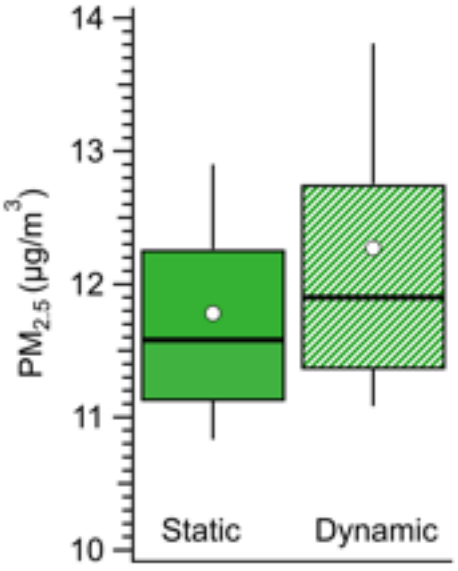

Figure S7: Boxplots for static and dynamic models when $\alpha = 12$ and $\beta = 18$ hours in Equations 4 and 5 of the main manuscript.

Bibliography:

1. Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Robinson, A. L., Presto, A. A., & Subramanian, R. (2020). Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. Aerosol Science and Technology, 54(2), 1–15. https://doi.org/10.1080/02786826.2019.1623863
2. Hornung, R. W.; Reed, L. D. Estimation of Average Concentration in the Presence of Nondetectable Values. null 1990, 5, 46–51, Publisher: Taylor & Francis.
3. Tekindal, M. A.; Erdoğan, B. D.; Yavuz, Y. Evaluating Left-Censored Data Through Substitution, Parametric, Semi-parametric, and Nonparametric Methods: A Simulation Study. Interdisciplinary sciences : computational life sciences 2015, 9, 153–172, ISBN:1913-2751.
4. Jain, S., Presto, A. A. & Zimmerman, N. Spatial Modeling of Daily PM2.5, NO2, and CO Concentrations Measured by a Low-Cost Sensor Network: Comparison of Linear, Machine Learning, and Hybrid Land Use Models. Environ. Sci. Technol. (2021) doi:10.1021/acs.est.1c02653.
5. Allegheny County GIS Group. Allegheny County Land Cover Areas. 2015. https://services1.arcgis.com/vdNDkVykv9vEWFX4/arcgis/rest/services/Land_Cover/FeatureServer