High-resolution geospatial database: national criteria-air-pollutant concentrations in the contiguous U.S., 2016 – 2020

Tianjun Lu[1]*, Sun-Young Kim[2], Julian D. Marshall[3]

[1]Department of Epidemiology and Environmental Health, College of Public Health, University of Kentucky, Lexington, Kentucky 40536, United States

[2]Department of Cancer AI and Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Gyeonggi-do, 10408, Korea

[3]Department of Civil & Environmental Engineering, University of Washington, 201 More Hall, Seattle, Washington 98195, United States

*Corresponding author; Telephone: 1-859-218-2229; Email: tianjunlu@uky.edu; Address: Department of Epidemiology and Environmental Health, College of Public Health, University of Kentucky, Lexington, Kentucky 40536, United States

Abstract

Concentrations estimates for ambient air pollution are used widely in fields such as environmental epidemiology, health impact assessment, urban planning, environmental equity, and sustainability. This study builds on previous efforts by developing an updated high-resolution geospatial database of population-weighted annual-average concentrations for six criteria air pollutants ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3) across the contiguous U.S. during a five-year period (2016-2020). We developed Land Use Regression (LUR) models within a partial-least-square – universal kriging framework by incorporating several land use, geospatial, and satellite – based predictor variables. The LUR models were validated using conventional and clustered cross-validation, with the former consistently showing superior performance in capturing the variability of air quality. Most models demonstrated reliable performance (e.g., mean squared error – based $R^2 > 0.8$, standardized root mean squared error < 0.1). We used the best modeling approach to develop estimates by Census Block, which were then population-weighted averaged at Census Block Group, Census Tract, and County geographies. Our database provides valuable insights into the dynamics of air pollution, with utility for environmental risk assessment, public health, policy, and urban planning.

Keywords

Criteria air pollutant, land use regression, empirical model, exposure assessment, geospatial data, environmental disparity, environmental hazard

Key findings

A high-resolution geospatial database of population-weighted annual-average concentrations for six criteria air pollutants ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3) across the contiguous U.S. during a five-year period (2016-2020) was created at Census Block Group, Census Tract, and County geographies. We developed Land Use Regression (LUR) models that incorporated a variety of land use, geospatial, and satellite – based predictor variables. Our database is publicly available and could inform analyses and policies in risk assessment, public health, environmental justice, and urban planning.

Dataset details

Identifier: https://www.caces.us/data.

Creator: Tianjun Lu, Sun-Young Kim, Julian D. Marshall

Dataset correspondence: jdmarsh@uw.edu

Title: National Census-block-level criteria-air-pollutant concentrations in the contiguous U.S., 2016 – 2020

Publisher: CACES

Publication year: 2024

(Resource type): web database

(Version): version 1.

1. Introduction

High spatial-resolution estimates for levels of ambient (outdoor) air pollution are useful for researchers and practitioners in epidemiology, public health and policy, health disparities, environmental equity, urban planning, and climate change (Bonell et al., 2023; Burger et al., 2024; Clark et al., 2014; Coleman et al., 2020; Hankey et al., 2017; Hystad et al., 2011; Nadal et al., 2015; Wang et al., 2023). Databases of ambient air pollution can provide spatially precise information for identifying and analyzing local pollution patterns, thereby allowing for more accurate risk assessment, air quality management, and locally-tailored interventions and policies (Allen & Barn, 2020; Bechle et al., 2023; Colmer et al., 2020; Lu et al., 2022; Lu et al., 2023; Zwack et al., 2011).

Empirical models of air pollution – also called land-use regression (LUR) or Integrated Empirical Geographic (IEG) models – are a geospatial approach that can predict, with fine spatial resolution, concentrations at unmeasured locations (Amini et al., 2017; Di et al., 2019a; Hankey & Marshall, 2015; Keller et al., 2015; Saha et al., 2021; Sampson et al., 2013; Van Donkelaar et al., 2019; Young et al., 2016). While details of the methods employed have evolved over time, LUR models exploit correlations between land use variables (e.g., the road network, population density, industry) and levels of ambient air pollution (Beelen et al., 2013; Hoek et al., 2008; Jerrett et al., 2005; Su et al., 2008).

Initially, LUR models were developed for single urban areas (Gilbert et al., 2005; Hankey et al., 2019; Henderson et al., 2007; Hoek et al., 2008; Jerrett et al., 2007; Lu et al., 2019; Marshall et al., 2008; Sanchez et al., 2018; Saraswat et al., 2013). Later, LUR was developed for national (Clark et al., 2014; Kerckhoffs et al., 2015; Kim et al., 2020; Novotny et al., 2011), continental (de Hoogh et al., 2016; Knibbs et al., 2016; Knibbs et al., 2018), and global scales

(Larkin et al., 2023; Larkin et al., 2017). Input variables have grown from land-use data, to now include satellite-derived pollution estimates (Bechle et al., 2015; Filonchyk & Peterson, 2023; Knibbs et al., 2018; Roy, 2021), predictions from mechanistic models (e.g., chemical transport models) (de Hoogh et al., 2016; Goldberg et al., 2019), and microscale features (e.g., street view and point of interest) (Lu et al., 2021; Messier et al., 2018; Qi et al., 2022; Qi & Hankey, 2021). Methods have shifted to better allow for non-criteria pollutants (Messier et al., 2018; Saha et al., 2022; Saraswat et al., 2013) and larger amounts of input data (i.e., more measured concentrations, more independent variables; "big data" techniques) and more-rapid, more-sophisticated mathematical modeling (e.g., statistical approach selection, variable prioritization, spatial coverage expansion; "machine learning") (Beckerman et al., 2013; Di et al., 2019b; Goldberg et al., 2019; Hoek et al., 2015; Keller et al., 2015; Kim et al., 2020; Lu et al., 2019; Lu et al., 2021; Masiol et al., 2019; US EPA, 2022; Van Donkelaar et al., 2019; Weichenthal et al., 2016; Wong et al., 2021).

A recent inter-comparison of spatial predictions by six national empirical models reported relatively close agreement among annual-average model predictions (Bechle et al., 2023). While the inputs and methods differed among models, the dependent data were consistent: publicly-available United States Environmental Protection Agency (US EPA) regulatory monitoring station data. That result suggests that, among the approaches they compared, the actual methods might not dramatically influence the resulting annual-average predictions, if the dependent variable is EPA regulatory air pollution measurements.

Here, we extend and update our previous modeling (Kim et al., 2020): national annual-average empirical models for criteria pollutants in the contiguous U.S. Specifically, in prior work we developed parsimonious national annual-average LUR models to estimate air pollutant concentrations for year-2015 and earlier. Here, we estimate concentrations for years-2016 to -2020, particularly allowing for ongoing health effect studies and exposure disparity research that would reveal evolving environmental impacts and facilitating the comparison of recent periods with historical data (Coleman et al., 2020; Lane et al., 2022; Liu et al., 2021). Prior estimates were publicly available for free download and have been integral to a range of applications (on average, currently receiving ~2 downloads per day); the same holds for the current estimates.

2. Data and Methods

In general, the three steps for empirical models are (1) model building, (2) model testing, and (3) model application. Below, we describe the dependent data (pollution concentrations), the independent data (i.e., "predictor variables", such as land uses and satellite data), and then the model-building, -testing, and -application stages. The approach employed here is consistent with prior work (i.e., years 2016-2020 here; years 2015 and earlier in prior work).

2.1. Dependent data: concentrations of criteria pollutants at regulatory monitoring stations

For regulatory purposes, such as ensuring compliance with national standards, EPA measures ambient concentrations of criteria air pollutants at regulatory monitoring stations throughout the U.S. Here, we employ data for year-2016 to -2020 for six criteria pollutants: nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and two types of particulate matter (PM_{10} [particles with diameters 10 micrometers or smaller] and $PM_{2.5}$ [particles with diameters 2.5 micrometers or smaller]). Data were obtained for all Air Quality System (AQS) monitoring locations via the EPA data repository (https://www.epa.gov/outdoor-air-quality-data).

The raw data are hourly for gaseous pollutants (NO_2, SO_2, CO, O_3) and daily for particulate matter ($PM_{2.5}$ and PM_{10}). We calculated annual-average concentrations of each pollutant (except O_3) at each site. Selection criteria, reflecting EPA criteria, are (1) a minimum of 18 hours of valid measurements per day, to be considered a valid day; (2) at least 244 valid days per year, or at least 61 days for monitors reporting every three days, or at least 41 days for those reporting every six days (i.e., at least 75% of intended measurements); (3) during the year, a maximum of 45 consecutive days without any measurements (Kim et al., 2020). In contrast, for O_3, the "annual average" metric includes only the heightened O_3 season from May through September, and represent the daily highest 8-hour average. Selection criteria for O_3 are (1) at least 18 hours of measurements to be considered a valid day and (2) at least 75% of valid days measured across the period. This process resulted in data collection from a range of 260 to 1,089 monitoring stations per pollutant strategically located across various US regions from 2016 to 2020. The breakdown by pollutant was as follows: $PM_{2.5}$ from 495 to 784 stations, PM_{10} from 290 to 454 stations, CO from 228 to 261 stations, NO_2 from 396 to 414 stations, SO_2 from 407 to 450 stations, and O_3 from 1,065 to 1,089 stations. Figure 1 illustrates the distribution of these monitoring stations. For LUR modeling, annual-average pollutant concentrations were square root transformed to adhere to the normality assumption.

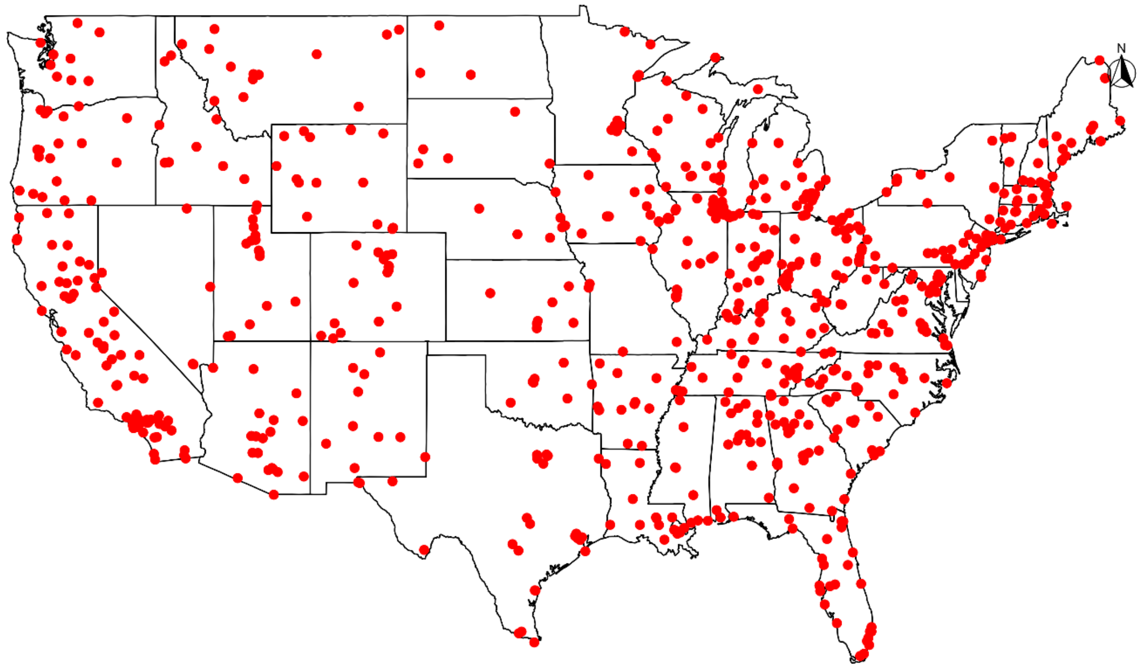

Figure 1. $PM_{2.5}$ monitoring stations in the contiguous US, 2016. Each red dot on the map represents a US EPA regulatory monitoring station.

2.2. Independent data: geographic and land use features and satellite estimates

We used the same predictor variables as in the prior study (Kim et al., 2020). Briefly, the variables cover ten categories: traffic, population, land use/land cover, sources, industrial emissions, vegetation, imperviousness, elevation, position, and satellite air pollution estimates (Table 1). These variables aim to capture various anthropogenic and natural factors that may impact air quality. The variables were quantified as count, length, and area within varying buffer sizes ranging from 50 meters to 15 kilometers; this approach aims to capture potential local and regional pollution impacts. Satellite-based pollution estimates (i.e., column abundance or surface estimates) are annual averages for NO_2, $PM_{2.5}$, SO_2, CO, and formaldehyde (HCHO), derived from multiple satellite products (e.g., aerosol optical depth [AOD], ozone monitoring instrument [OMI], and estimates from chemical transport models) (Boersma et al., 2011; Deeter et al., 2017). Consistent with Kim et al. (2020), the HCHO satellite data are a 12-year average (2005 to

2016) rather than annual averages. The resolution of the gridded annual satellite-based estimates varied (NO_2, $PM_{2.5}$, HCHO: 0.1°×0.1°; SO_2, CO: 0.25°×0.25°). Overall, we employ a total of ~360 variables. Additional details regarding data processing and variable calculation are in (Kim et al., 2020).

Table 1. List of geographic and land use variables and satellite estimates used in the LUR models

Category	**Measure**	**Spatial resolution**	**Description**	**Data source[a]**
Traffic	Length/density in buffer (km)	0.05-15 km	Any road, truck route, intersections, etc.	TeleAtlas
Population	Count in buffer	Block group	Population in block groups (0.5-3 km)	U.S. Census
Land use/land cover	Area in buffer (%)	30 m	Built land, open space, agricultural land, etc. (0.05-15 km)	NLCD
Sources	Length in buffer (m)	Point	Distance to the nearest source (e.g., railroad, airport)	NEI
Emissions	Point in buffer (ton)	Point	Sum of site-specific facility emissions (3-30 km)	NEI
Vegetation	Area in buffer	30 m	Normalized difference vegetation index (0.5-10 km)	University of Maryland
Impervious	Area in buffer (%)	30 m	Impervious surface value (0.05-5 km)	NLCD
Elevation	Value	30 m	Elevation above sea levels; counts of points above or below a threshold (1-5 km)	USGS NED
Position	Coordinates	Point	Latitude, longitude	EPA
Satellite estimates	Column abundance or surface ($\mu g/m^3$ or ppb)	10-25 km	Satellite-based air pollution estimates (NO_2, SO_2, CO, HCHO, $PM_{2.5}$)	Multiple sources[a]

Note: [a]Satellite-derived annual average estimates for $PM_{2.5}$ (1998-2014, 0.1° * 0.1° grid), NO_2 (2004-2015; 0.1° * 0.1° grid), SO_2 (2005-2016; 0.25° * 0.25° grid), and CO (2001-2016; 0.25° * 0.25° grid). Detailed information about the data sources can be found in Kim et al., 2020 and Lu et al., 2021. TeleAtlas provides dynamic location-based data. NLCD is National Land Cover Database; NEI is National Emission Inventory; USGS NED is National Elevation Dataset (NED) of the U.S. Geological Survey (USGS).

2.3. Modeling -building and -testing

Consistent with previous modeling approaches (Kim et al., 2020; Sampson et al., 2013; Young et al., 2016), we employed a hybrid modeling approach that integrates partial least squares (PLS) with universal kriging (PLS-UK). Specifically, the two components (1) address variance through kriging that utilizes an exponential covariance function for the variogram and (2) managing the mean component via PLS that reduces the dimensions of predictor variables in the linear regression process. This combination not only encompasses a broad range of geographic factors but also mitigates the likelihood of generating extreme predictions (Mehmood et al., 2012). Covariance parameters in kriging and regression parameters for PLS summary predictors were selected based on a maximum likelihood approach. We adopted two and three PLS predictors based on cross-validation, respectively, following previous studies using a similar modeling framework (Kim et al., 2020; Young et al., 2016).

To further investigate the model-building process, we employed forward selection to choose an optimal subset of variables from the full set, for use in the PLS approach. That is, the process begins by identifying the variable with the highest correlation to the dependent variable and incrementally adding variables that yield the highest partial F-statistic in combination with previously selected variables. This process examines how the number of selected variables influences the model performance.

For model testing, we executed two types of 10-fold cross-validation (CV) training and testing: (1) in conventional CV, monitoring sites are divided into 10 groups randomly; (2) in spatially clustered CV, k-means is used to establish 10 spatial groups (clusters) spatially (Young et al., 2016). In both CV, there are 10 iterations, each involving setting aside one group for model-testing and using the remaining groups for model-building. Conventional CV evaluates model performance at random locations; spatially clustered CV reflects model performance at sites distant from monitoring stations.

To evaluate model performance via cross-validation comparisons, we utilized standardized root-mean-square error (sRMSE) and a mean squared error (MSE) – based R^2 metric. sRMSE, calculated as RMSE divided by the mean concentrations of all monitors, allows for comparisons across pollutants. MSE-R^2 (hereafter R^2), defined as one minus the ratio of the sum of squared prediction errors to the sum of squared deviations from the observation mean, assesses the fit of the predictions to the 1:1 line rather than the regression line (Knibbs et al., 2016; Wang et al., 2012).

2.4. Model application

The most effective LUR models identified through this process were then applied to our nationwide model predictions for each pollutant and year. After we identified the best models for each pollutant and respective year, we predicted the annual average concentrations for approximately 7 million Census Block centroids with nonzero population throughout the contiguous U.S. Subsequently, we calculated population-weighted averages at several geographical levels, including Census Block Groups, Census Tracts, and Counties for years 2016 – 2020, aligning with the 2010 Census boundaries. Of the 30 models attempted (i.e., 6 pollutants, 5 years), we were unable to generate one of the models [O3, 2020] because of data constraints, leaving 29 resulting pollutant-year models. This approach was designed to offer our estimates across diverse spatial resolutions for all criteria pollutants, catering to a variety of application needs. All LUR model development were carried out in R (version 4.2.1).

3. Results and discussion

3.1. Monitoring data

The main trend for monitoring data we investigated (Figure 2) is, with the exception of O_3, a slight decline in air pollutant concentrations. This pattern is consistent with the previous years of regulatory monitoring data, indicating improved air quality measures or other environmental policy impacts nationwide (Kim et al., 2020). For $PM_{2.5}$ and PM_{10}, the concentrations slight increased from 2016 to 2018, but decreased in 2019 and 2020. NO_2 was steady from 2016 to 2019 and showed a slight decrease in concentrations in 2020. SO_2 generally demonstrated a downward trend in mean concentrations and consistent year-on-year reduction while CO and O_3 displayed minor yearly fluctuations with no apparent trend in concentrations.

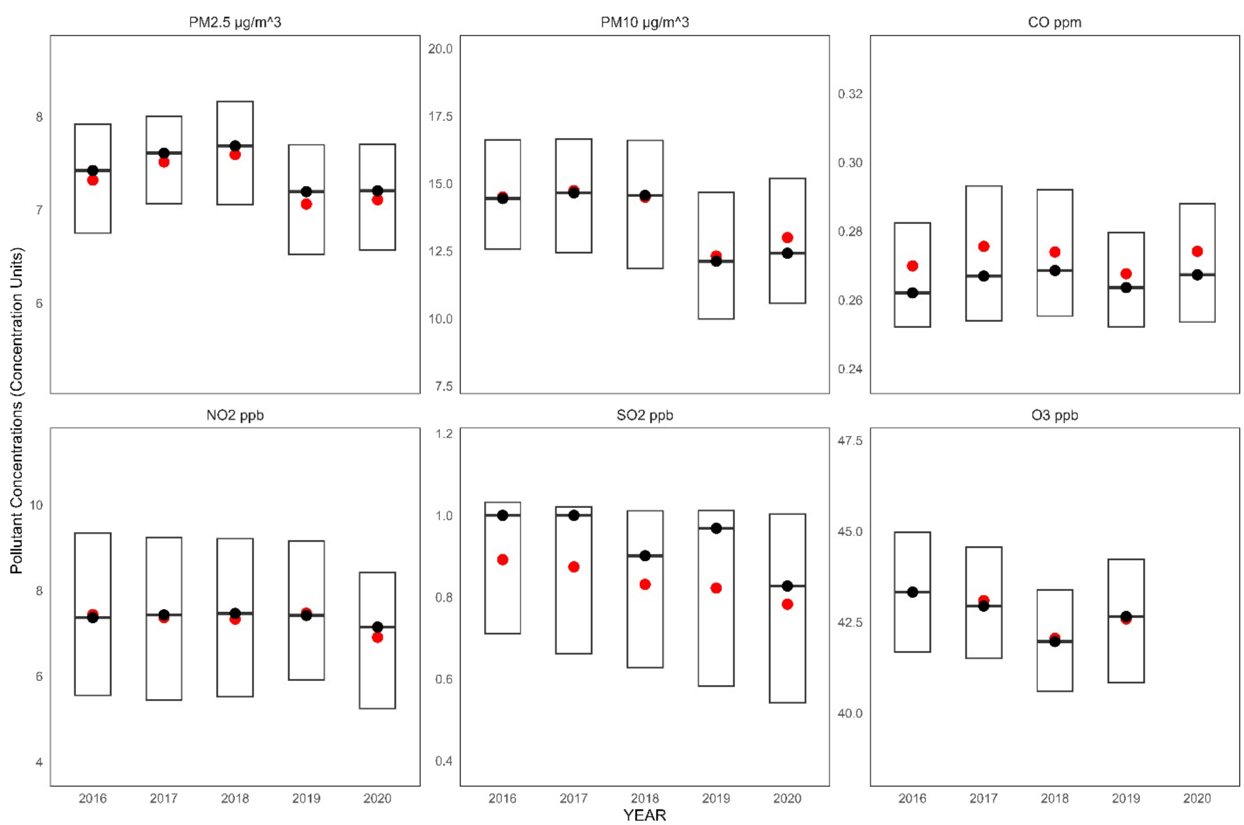

Figure 2. Boxplots of annual average concentrations of six criteria air pollutants across regulatory monitoring sites for 2016 – 2020 in the contiguous U.S. Each box plot displays the median concentration (black dot) and mean concentration (red dot) and the interquartile range

(IQR, represented by the box) of pollutant concentrations for each year. The unit of $PM_{2.5}$ and PM_{10} is μg/m^3. The unit of CO is ppm. The unit of NO_2, SO_2, and O_3 is ppb.

3.2. Model evaluation

Overall, for each pollutant and across both CV methodologies, LUR models constructed with around 30 predictor variables consistently yielded better R^2 and lower sRMSE as compared to models with more predictors (Figure 3). This pattern underscores the diminishing returns and potential overfitting when too many variables are included (Kerckhoffs et al., 2019; Li et al., 2017). The temporal stability from 2016 to 2020 of model performance also highlights the robustness of the LUR models. These results imply that a more parsimonious approach to model development, i.e., selecting only the most pertinent categories of predictor variables, can provide more reliable models; that finding is consistent with prior research (Kim et al., 2020; Meng et al., 2015).

In terms of the two CV approaches, LUR models validated using the conventional CV method consistently outperformed those using the spatially clustered CV approach, as indicated by higher R^2 and lower sRMSE values (Figure 3). That result is consistent with expectations and with the previous modeling effort (Kim et al., 2020). As mentioned above, clustered CV reflect model performance far from other monitors.

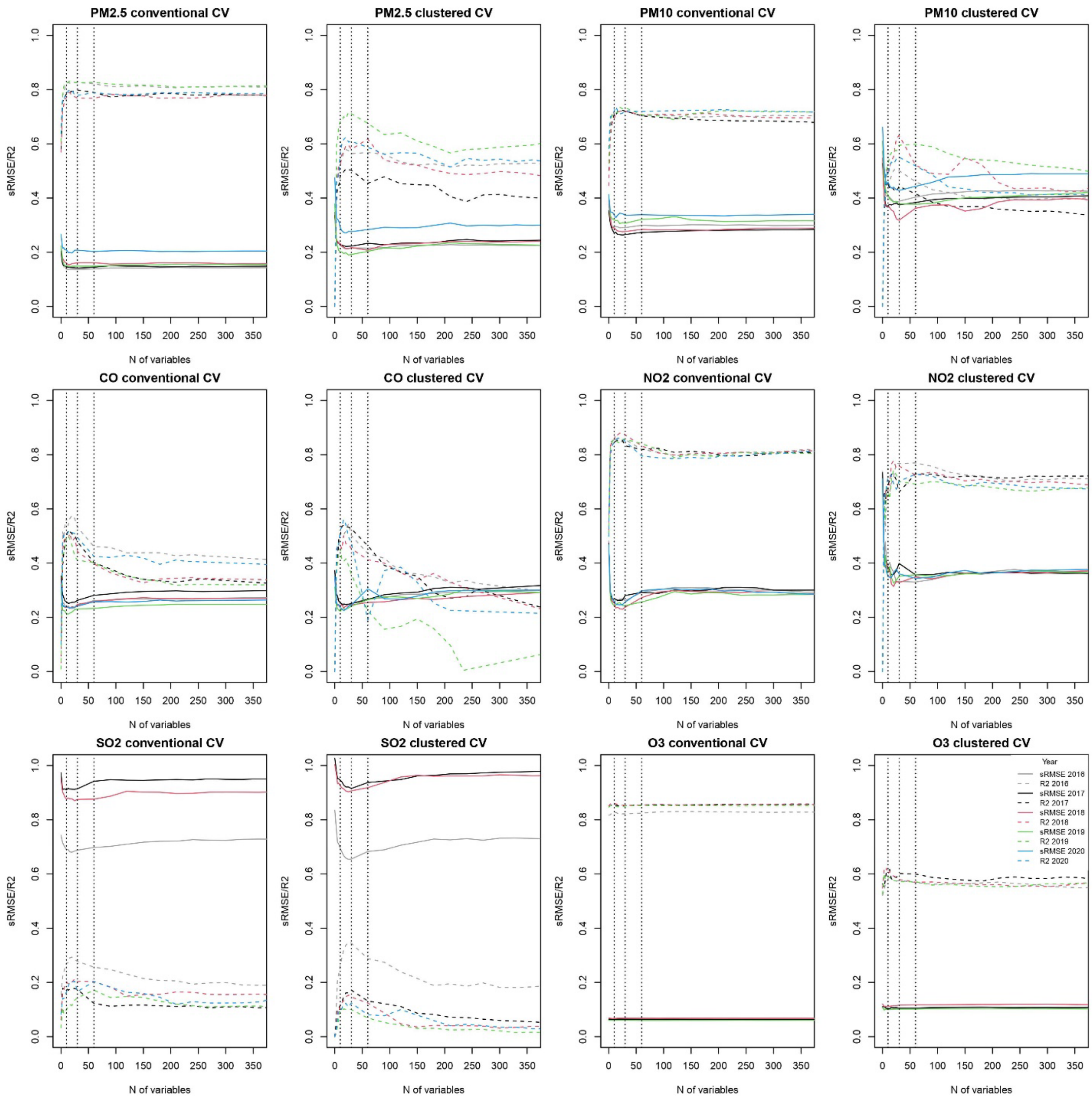

Figure 3. sRMSEs and R^2s of the conventional and clustered CV of LUR models for six criteria air pollutants, 2016 – 2020, categorized by the number of predictor variables. The solid lines represent sRMSEs, and the dashed lines represent R^2s.

Table 2 shows the performance of LUR models using the "best" number of predictor variables. In general, model performance is consistently higher for $PM_{2.5}$, PM_{10}, NO_2, and O_3 than for CO and SO_2, particularly when using conventional CV. Over the five-year span, this trend is evident through the sustained R^2 values above 0.70 for $PM_{2.5}$ and PM_{10} and exceeding 0.80 for NO_2 and O_3, underscoring the robustness of the LUR models for these pollutants. Scatterplots of the LUR models using the "best" number of predictor variables for six criteria air pollutants in 2016 (Figure 4) illustrate the level of agreement between predicted versus measured air pollution levels. Scatterplots reveal similar patterns as in Table 2.

Table 2. sRMSEs and R^2s of the conventional and clustered CV of LUR models using the "best" number of predictor variables for six criteria air pollutants, 2016 – 2020

		Conventional CV		Clustered CV	
Pollutant	**Year**	**R^2**	**sRMSE**	**R^2**	**sRMSE**
$PM_{2.5}$	2016	0.83	0.14	0.56	0.22
$PM_{2.5}$	2017	0.80	0.14	0.5	0.22
$PM_{2.5}$	2018	0.77	0.16	0.58	0.22
$PM_{2.5}$	2019	0.83	0.15	0.71	0.19
$PM_{2.5}$	2020	0.78	0.21	0.61	0.28
PM_{10}	2016	0.72	0.29	0.50	0.39
PM_{10}	2017	0.72	0.27	0.44	0.37
PM_{10}	2018	0.72	0.28	0.63	0.32
PM_{10}	2019	0.73	0.31	0.59	0.38
PM_{10}	2020	0.72	0.34	0.55	0.43
CO	2016	0.53	0.24	0.50	0.25
CO	2017	0.48	0.26	0.53	0.25
CO	2018	0.44	0.25	0.45	0.25
CO	2019	0.41	0.23	0.36	0.24
CO	2020	0.49	0.24	0.48	0.25
NO_2	2016	0.86	0.26	0.77	0.33
NO_2	2017	0.83	0.28	0.66	0.40
NO_2	2018	0.87	0.24	0.76	0.33
NO_2	2019	0.86	0.24	0.70	0.35
NO_2	2020	0.85	0.26	0.70	0.36
SO_2	2016	0.28	0.69	0.34	0.66
SO_2	2017	0.17	0.91	0.17	0.92
SO_2	2018	0.20	0.88	0.15	0.91
SO_2	2019	0.14	1.65	0.10	1.69
SO_2	2020	0.17	1.45	0.12	1.49
O_3	2016	0.82	0.07	0.58	0.10
O_3	2017	0.85	0.06	0.60	0.11
O_3	2018	0.85	0.07	0.58	0.12
O_3	2019	0.85	0.06	0.57	0.10

Note: sRMSE, calculated as RMSE divided by the mean concentrations of all monitors. R^2 represents MSE-R^2, defined as one minus the ratio of the sum of squared prediction errors to the sum of squared deviations from the observation mean, and it assesses the fit of the predictions to the 1:1 line rather than the regression line. In conventional CV, monitoring sites are divided into 10 groups randomly; in spatially clustered CV, k-means is used to establish 10 spatial groups (clusters) spatially (Young et al., 2016).

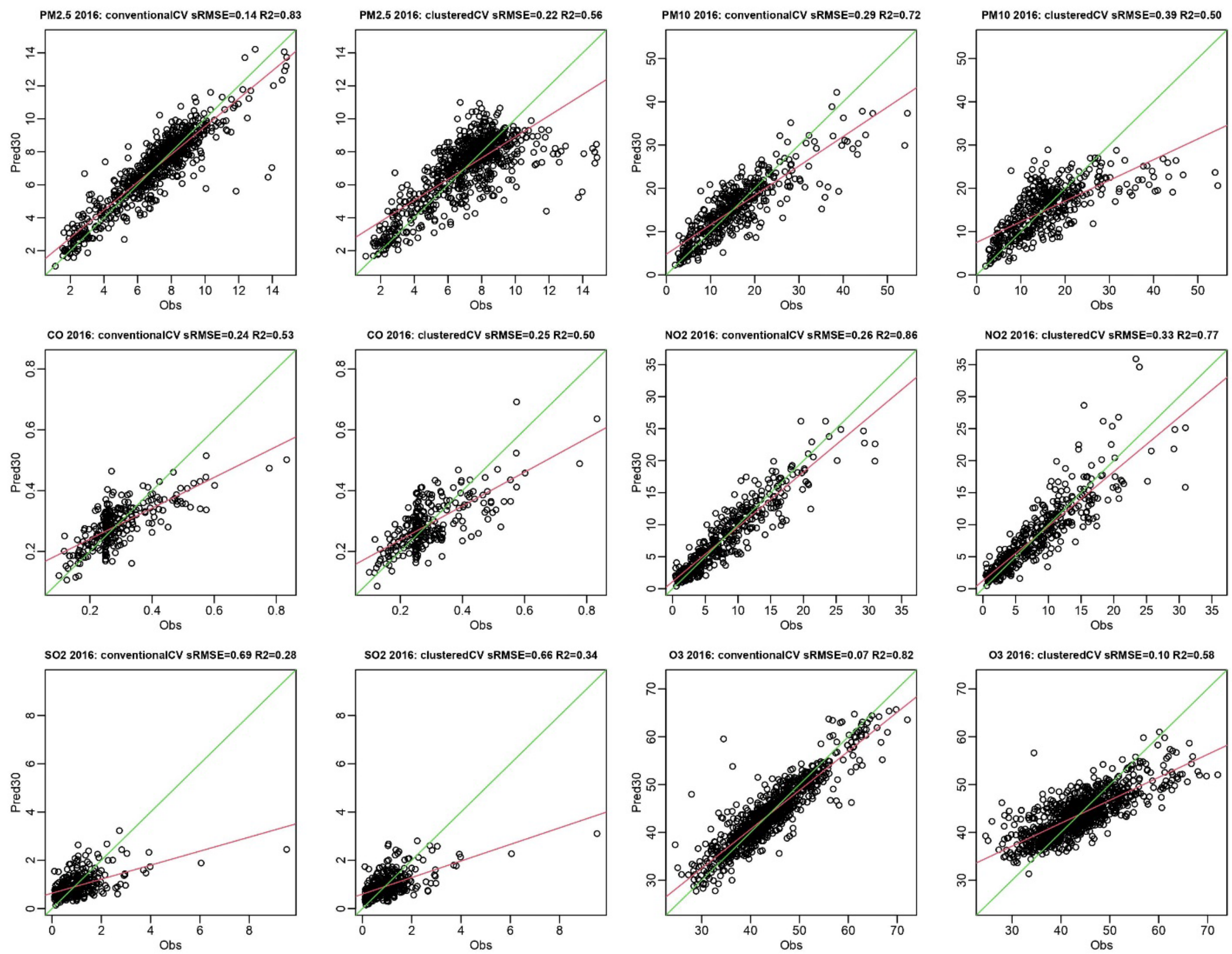

Figure 4. Scatterplots of the sRMSEs and R^2s of the conventional and clustered CV of LUR models using the "best" number of predictor variables for six criteria air pollutants in 2016. Green line indicates the 1:1 line and red line indicates the fitted line.

3.3. Model application

After model-building and -testing, we applied the model across geographies, estimating concentrations at each Census Block and then population-weighted averaging to the Block Group (year-2016: Figure 5), tract, and county. The maps revealed the heterogeneous distribution of each air pollutant. As expected, the $PM_{2.5}$ concentrations were notably higher in the California and along some urbanized areas of the East Coast, reflecting the combined impact of energy production and dense population centers (Kelly et al., 2021). PM_{10} levels showed similar patterns likely also reflecting the influence of both natural desert dust and anthropogenic sources (Chen et al., 2010). SO_2 levels showed slightly higher levels in Southern and Central states, potentially near industrial sources (Nopmongcol et al., 2019). O_3 concentrations demonstrated a broad distribution, with higher levels in the Southern and Southwestern states, which could be attributed to these regions' warmer climates, conducive to O_3 formation (Jiao et al., 2016). These maps underscore the diverse spatial patterns of air pollutant concentrations across the U.S., reflecting regional differences in emission sources, environmental policies, and natural factors.

The summary statistics for population-weighted annual average concentrations of air pollutants from 2016 to 2020 revealed an upward trend in $PM_{2.5}$ exposure, with both median and mean concentrations rising over the five-year period (Table 3). In contrast, NO_2 and SO_2 levels decreased, suggesting improvements in these pollutants nationwide. PM_{10} and O_3 concentrations fluctuated year-to-year without a clear long-term trend, indicating variability in exposure to these pollutants across the U.S. population.

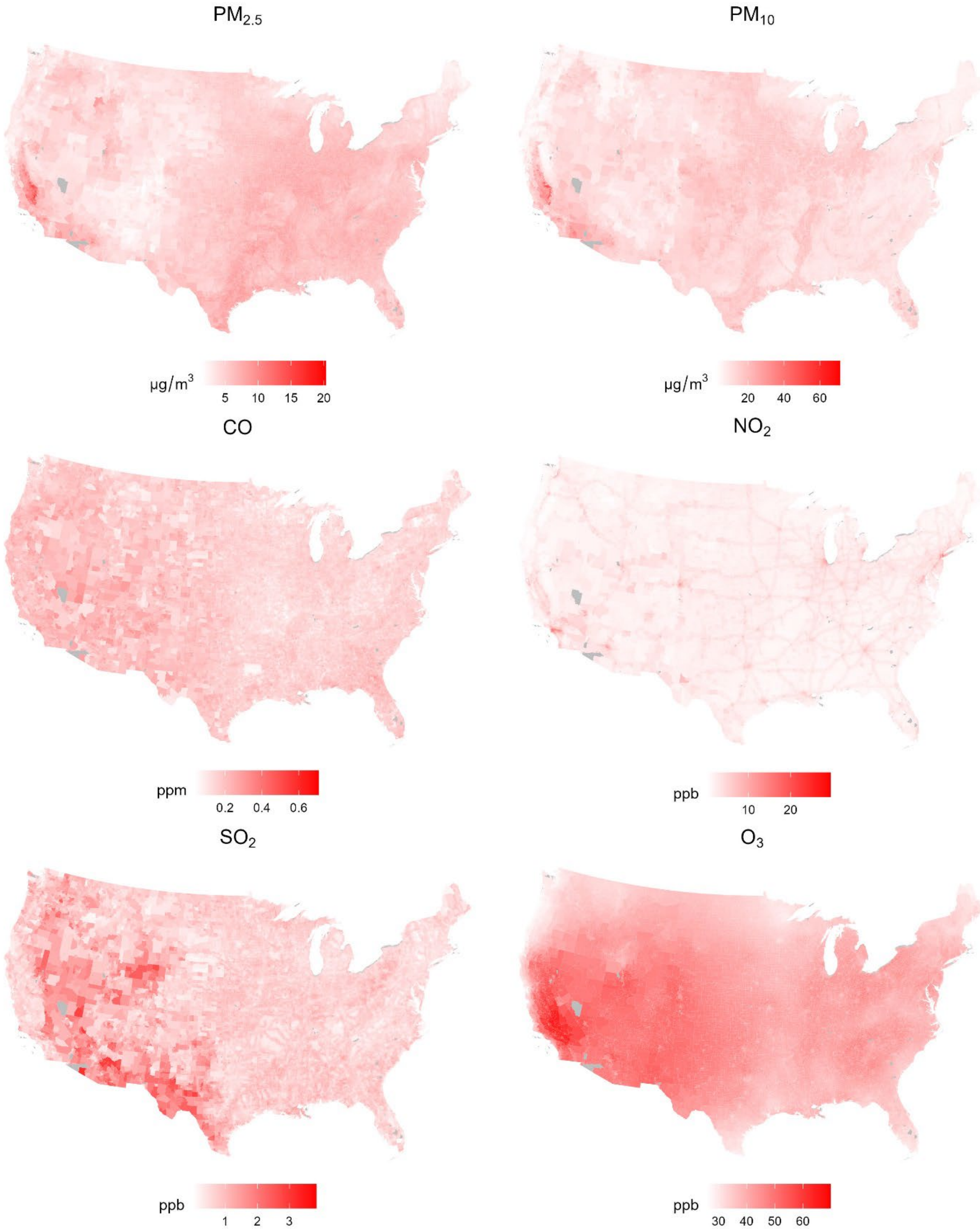

Figure 5. Maps of Census Block Group population-weighted annual average concentrations for six criteria air pollutants in 2016 based on the "best" LUR models.

Table 3. Summary statistics[a] of population-weighted annual average concentrations[b] across all Census Block Groups for the contiguous U.S., 2016-2020, based on the "best" LUR models

Pollutant	Year	P10	P25	Median	P75	P90	Mean	SD
$PM_{2.5}$	2016	5.49	6.43	7.49	8.27	9.07	7.40	1.56
$PM_{2.5}$	2017	5.75	6.68	7.61	8.48	9.39	7.64	1.60
$PM_{2.5}$	2018	5.80	6.69	7.70	8.72	10.05	7.86	1.82
$PM_{2.5}$	2019	5.14	6.15	7.18	8.03	8.93	7.10	1.55
$PM_{2.5}$	2020	5.27	6.17	7.19	8.33	10.00	7.55	2.37
PM_{10}	2016	10.51	12.94	16.06	19.38	22.92	16.70	5.63
PM_{10}	2017	10.64	13.03	15.82	19.44	22.99	16.71	5.65
PM_{10}	2018	10.71	13.03	15.99	19.87	24.07	16.93	5.66
PM_{10}	2019	9.33	11.71	14.54	17.91	21.05	15.09	4.99
PM_{10}	2020	9.02	12.01	15.95	20.86	26.37	16.97	7.04
CO	2016	0.16	0.19	0.23	0.27	0.31	0.23	0.06
CO	2017	0.17	0.19	0.22	0.27	0.32	0.23	0.07
CO	2018	0.17	0.19	0.23	0.27	0.31	0.23	0.06
CO	2019	0.15	0.19	0.22	0.25	0.29	0.22	0.05
CO	2020	0.16	0.19	0.22	0.26	0.30	0.23	0.06
NO_2	2016	2.31	3.61	5.97	8.71	12.72	6.81	4.23
NO_2	2017	2.22	3.24	5.33	8.22	12.49	6.43	4.31
NO_2	2018	2.23	3.43	5.54	8.07	11.93	6.38	3.99
NO_2	2019	2.29	3.49	5.66	8.27	11.80	6.43	3.85
NO_2	2020	1.94	3.19	5.16	7.57	10.91	5.83	3.57
SO_2	2016	0.48	0.60	0.76	0.94	1.13	0.79	0.28
SO_2	2017	0.38	0.52	0.68	0.85	1.04	0.70	0.28
SO_2	2018	0.34	0.47	0.62	0.80	0.99	0.65	0.27
SO_2	2019	0.26	0.40	0.59	0.82	1.04	0.63	0.32
SO_2	2020	0.24	0.38	0.55	0.73	0.91	0.57	0.27
O_3	2016	34.81	39.22	42.48	44.64	48.08	42.12	5.69
O_3	2017	35.24	39.00	41.88	44.46	50.05	42.24	6.16
O_3	2018	34.66	37.81	40.75	42.73	49.83	41.15	5.94
O_3	2019	35.03	38.35	41.55	44.08	48.71	41.61	5.68

Note: [a]P10 represents the 10th percentile. P25 represents the 25th percentile. P75 represents the 75th percentile. P90 represents the 90th percentile. SD represents standard deviation.
[b]The unit on $PM_{2.5}$ and PM_{10} is μg/m^3. The unit on CO is ppm. The unit on NO_2, SO_2, and O_3 is ppb.

3.4. Comparison with the previously developed database (1979 – 2015)

The consistent methodology across the current and historical datasets enables users to track long-term trends in air pollution exposure with confidence. Notably, the trend of decreasing NO_2 and SO_2 levels aligns with historical data, suggesting ongoing improvements in air quality. However, the recent increase in $PM_{2.5}$ and PM_{10} concentrations as observed in our updated models disagreed with the previous models' decreasing trend. This finding is consistent with some recent studies (Bekbulat et al., 2021), which warrants further investigation into the sources of particulate pollution. The variability observed in the O_3 concentrations is consistent with the historical dataset (Kim et al., 2020). The prediction maps further emphasized the spatial consistency in air pollution distribution as compared to previous periods. For example, areas historically impacted by higher concentrations of $PM_{2.5}$ and PM_{10}, such as California, continue to show elevated levels, while regions with industrial activities remain as hotspots for SO_2. These spatial and temporal trends of consistency highlight the persistent influence of both anthropogenic activities and natural impacts on ambient air quality.

3.5. Dataset location and format

Model predictions can be freely downloaded online at https://www.caces.us/data. The available dataset provides population-weighted annual-average concentrations for the six criteria air pollutants at three geographic levels (Census Block Groups; Census Tracts; Counties) during 2016 – 2020 (as noted above, O_3 estimates are for summer months, and estimates for 2020 are unavailable).

3.6. Limitations and future work

Our comprehensive high-resolution geospatial database supplements previously developed air pollution database (1979 – 2015) and provides detailed estimates of population-weighted annual-average concentrations for the six criteria pollutants ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3) across varying geographic scales, spanning the combined years 1979 – 2020.

While our model provides comprehensive geospatial estimates of air pollutant concentrations, it is not without limitations. The reliance on regulatory monitoring data can lead to underrepresentation of air quality in regions with sparse monitoring infrastructure, potentially affecting the precision of estimates in rural or remote areas (deSouza & Kinney, 2021; Kelp et al., 2022). Future enhancements could include the integration of other remote sensing products (e.g., Google Earth Engine) and low-cost sensor networks (e.g., PurpleAir) to increase spatial coverage, especially in under-monitored regions (Brook et al., 2018; Lee et al., 2017; Lu et al., 2022). Modeling performance is not as strong for CO and SO_2 as for other pollutants. That result may reflect several possible aspects, including the number and placement of monitors, spatial patterns, and the degree to which those patterns correlate with land-use or other data used in our model. Future research might incorporate more detailed emission inventory data into the model development, aiming to capture point sources and industrial emissions that could significantly affect local concentrations. Our current modeling framework is PLS-UK; future modeling approaches could include other geospatial and empirical models (e.g., machine learning, geographically weighted regression) (Lloyd et al., 2023; Lotfata, 2022; Lu et al., 2021). Furthermore, the temporal resolution here is confined to annual averages, which incorporates but does not shed light on short-term variability, episodic events, or seasonal patterns (Dons et al., 2014); future work could advance the temporal resolution of our models to capture temporal dynamics; this approach may, potentially, inform timely and targeted air quality interventions

(Gladson et al., 2022; Masiol et al., 2018; Yuan et al., 2022). To be consistent with the prior modeling efforts (Kim et al., 2020), our approach is based on regulatory monitors and so is confined to criteria pollutants; future work could consider air toxics or emerging pollutants, if monitoring data were available (Amini et al., 2017; Jedynska et al., 2014; Saha et al., 2021).

4. Conclusion

We developed a high-resolution geospatial database consisting of the population-weighted annual average concentrations of six key criteria air pollutants across the contiguous U.S. from 2016 to 2020, at several geographical levels, including Census Block Groups, Census Tracts, and Counties. Our models were developed using a PLS-UK framework; the models achieved reasonable performance, consistent with our earlier published models.

Acknowledgements

This publication was developed as part of the Center for Air, Climate, and Energy Solutions (CACES), which was supported under Assistance Agreement No. R835873 awarded by the U.S. Environmental Protection Agency (EPA) for an Air, Climate, and Energy (ACE) center. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. We thank Amanda J. Gassett and Logan Piepmeier from the University of Washington for the assistance in generating the model variables. We also thank Dr. Steve Hankey from Virginia Tech for valuable comments on the model development.

CrediT Authorship Contribution Statement

Tianjun Lu: Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. Sun-Young Kim: Data curation, Methodology, Writing – review & editing. Julian D. Marshall: Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Conflict of Interest Statement

All authors have no conflict of interest to declare.

Data Availability Statement

The high-resolution geographic database (estimated ambient annual-average concentrations) can be downloaded as CSV files freely from the official website of the Center for Air, Climate, and Energy Solutions (CACES): https://www.caces.us/data.

References

Allen, R. W., & Barn, P. (2020). Individual-and household-level interventions to reduce air pollution exposures and health risks: a review of the recent literature. *Current environmental health reports*, *7*, 424-440.

Amini, H., Yunesian, M., Hosseini, V., Schindler, C., Henderson, S. B., & Künzli, N. (2017). A systematic review of land use regression models for volatile organic compounds. *Atmospheric environment*, *171*, 1-16.

Bechle, M. J., Bell, M. L., Goldberg, D. L., Hankey, S., Lu, T., Presto, A. A., Robinson, A. L., Schwartz, J., Shi, L., & Zhang, Y. (2023). Intercomparison of six national empirical models for PM2. 5 air pollution in the contiguous US. *Findings*(November). https://doi.org/https://doi.org/10.32866/001c.89423

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2015). National spatiotemporal exposure surface for NO2: monthly scaling of a satellite-derived land-use regression, 2000–2010. *Environmental science & technology*, *49*(20), 12297-12305.

Beckerman, B. S., Jerrett, M., Serre, M., Martin, R. V., Lee, S.-J., Van Donkelaar, A., Ross, Z., Su, J., & Burnett, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM2. 5 in the contiguous United States. *Environmental science & technology*, *47*(13), 7233-7241.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., & Marcon, A. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe–The ESCAPE project. *Atmospheric environment*, *72*, 10-23.

Bekbulat, B., Apte, J. S., Millet, D. B., Robinson, A. L., Wells, K. C., Presto, A. A., & Marshall, J. D. (2021). Changes in criteria air pollution levels in the US before, during, and after Covid-19 stay-at-home orders: Evidence from regulatory monitors. *Science of the total environment*, *769*, 144693.

Boersma, K., Eskes, H., Dirksen, R., Van Der A, R., Veefkind, J., Stammes, P., Huijnen, V., Kleipool, Q., Sneep, M., & Claas, J. (2011). An improved tropospheric NO 2 column retrieval algorithm for the Ozone Monitoring Instrument. *Atmospheric Measurement Techniques*, *4*(9), 1905-1928.

Bonell, A., Part, C., Okomo, U., Cole, R., Hajat, S., Kovats, S., Sferruzzi-Perri, A. N., & Hirst, J. E. (2023). An expert review of environmental heat exposure and stillbirth in the face of climate change: Clinical implications and priority issues. *BJOG: An International Journal of Obstetrics & Gynaecology*.

Brook, J. R., Setton, E. M., Seed, E., Shooshtari, M., Doiron, D., Awadalla, P., Brauer, M., Hu, H., McGrail, K., Stieb, D., Subarrao, P., Demers, P., Manuel, D., McLaughlin, J., Carlsten, C., Azad, M., Atkinson, S., Burnett, R., Lou, W., . . . Consortium, C. T. C. U. E. H. R. (2018). The Canadian Urban Environmental Health Research Consortium – a protocol for building a national environmental exposure data platform for integrated analyses of urban form and health. *BMC Public Health*, *18*(1), 114. https://doi.org/10.1186/s12889-017-5001-5

Burger, M., Gubler, M., & Brönnimann, S. (2024). High-resolution dataset of nocturnal air temperatures in Bern, Switzerland (2007–2022). *Geoscience Data Journal*(January). https://doi.org/https://doi.org/10.1002/gdj3.237

Chen, J., Ying, Q., & Kleeman, M. J. (2010). Source apportionment of wintertime secondary organic aerosol during the California regional PM10/PM2.5 air quality study.

Atmospheric environment, *44*(10), 1331-1340. https://doi.org/https://doi.org/10.1016/j.atmosenv.2009.07.010
Clark, L. P., Millet, D. B., & Marshall, J. D. (2014). National patterns in environmental injustice and inequality: outdoor NO2 air pollution in the United States. *PloS one*, *9*(4), e94431.
Coleman, N. C., Burnett, R. T., Ezzati, M., Marshall, J. D., Robinson, A. L., & Pope III, C. A. (2020). Fine particulate matter exposure and cancer incidence: analysis of SEER cancer registry data from 1992–2016. *Environmental health perspectives*, *128*(10), 107004.
Colmer, J., Hardman, I., Shimshack, J., & Voorheis, J. (2020). Disparities in PM2. 5 air pollution in the United States. *Science*, *369*(6503), 575-578.
de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R. V., Marshall, J. D., Bechle, M. J., Cesaroni, G., Pradas, M. C., Dedele, A., & Eeftens, M. (2016). Development of West-European PM2. 5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental research*, *151*, 1-10.
Deeter, M. N., Edwards, D. P., Francis, G. L., Gille, J. C., Martínez-Alonso, S., Worden, H. M., & Sweeney, C. (2017). A climate-scale satellite record for carbon monoxide: the MOPITT Version 7 product. *Atmospheric Measurement Techniques*, *10*(7), 2533-2555.
deSouza, P., & Kinney, P. L. (2021). On the distribution of low-cost PM2.5 sensors in the US: demographic and air quality associations. *Journal of Exposure Science & Environmental Epidemiology*, *31*(3), 514-524. https://doi.org/10.1038/s41370-021-00328-2
Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., & Lyapustin, A. (2019a). Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. *Environmental science & technology*, *54*(3), 1372-1384.
Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., & Lyapustin, A. (2019b). An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, *130*, 104909.
Dons, E., Van Poppel, M., Int Panis, L., De Prins, S., Berghmans, P., Koppen, G., & Matheeussen, C. (2014). Land use regression models as a tool for short, medium and long term exposure to traffic related air pollution. *Science of the total environment*, *476-477*, 378-386. https://doi.org/https://doi.org/10.1016/j.scitotenv.2014.01.025
Filonchyk, M., & Peterson, M. P. (2023). An integrated analysis of air pollution from US coal-fired power plants. *Geoscience Frontiers*, *14*(2), 101498.
Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. *Journal of the Air & Waste Management Association*, *55*(8), 1059-1063.
Gladson, L., Garcia, N., Bi, J., Liu, Y., Lee, H. J., & Cromar, K. (2022). Evaluating the utility of high-resolution spatiotemporal air pollution data in estimating local PM2. 5 exposures in California from 2015–2018. *Atmosphere*, *13*(1), 85.
Goldberg, D. L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z., & Streets, D. G. (2019). Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM2. 5 concentrations at 1 km resolution in the Eastern United States. *Atmospheric environment*, *199*, 443-452.
Hankey, S., Lindsey, G., & Marshall, J. D. (2017). Population-level exposure to particulate air pollution during active travel: planning for low-exposure, health-promoting cities. *Environmental health perspectives*, *125*(4), 527-534.

Hankey, S., & Marshall, J. D. (2015). Land use regression models of on-road particulate air pollution (particle number, black carbon, PM2. 5, particle size) using mobile monitoring. *Environmental science & technology*, *49*(15), 9194-9202.
Hankey, S., Sforza, P., & Pierson, M. (2019). Using mobile monitoring to develop hourly empirical models of particulate air pollution in a rural Appalachian community. *Environmental science & technology*, *53*(8), 4305-4315.
Henderson, S. B., Beckerman, B., Jerrett, M., & Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental science & technology*, *41*(7), 2422-2428.
Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, *42*(33), 7561-7578.
Hoek, G., Eeftens, M., Beelen, R., Fischer, P., Brunekreef, B., Boersma, K. F., & Veefkind, P. (2015). Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmospheric environment*, *105*, 173-180.
Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., van Donkelaar, A., Lamsal, L., Martin, R., & Jerrett, M. (2011). Creating national air pollution models for population exposure assessment in Canada. *Environmental health perspectives*, *119*(8), 1123-1129.
Jedynska, A., Hoek, G., Wang, M., Eeftens, M., Cyrys, J., Keuken, M., Ampe, C., Beelen, R., Cesaroni, G., Forastiere, F., Cirach, M., de Hoogh, K., De Nazelle, A., Nystad, W., Declercq, C., Eriksen, K. T., Dimakopoulou, K., Lanki, T., Meliefste, K., . . . Kooter, I. M. (2014). Development of land use regression models for elemental, organic carbon, PAH, and hopanes/steranes in 10 ESCAPE/TRANSPHORM European study areas. *Environmental science & technology*, *48*(24), 14435-14444. https://doi.org/10.1021/es502568z
Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, *15*(2), 185-204.
Jerrett, M., Arain, M., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N., Brook, J., Finkelstein, N., & Finkelstein, M. (2007). Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada. *Journal of Toxicology and Environmental Health, Part A*, *70*(3-4), 200-212.
Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S., & Buckley, K. (2016). Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmos Meas Tech*, *9*(11), 5281-5292. https://doi.org/10.5194/amt-9-5281-2016
Keller, J. P., Olives, C., Kim, S.-Y., Sheppard, L., Sampson, P. D., Szpiro, A. A., Oron, A. P., Lindström, J., Vedal, S., & Kaufman, J. D. (2015). A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environmental health perspectives*, *123*(4), 301-309.
Kelly, J. T., Jang, C., Timin, B., Di, Q., Schwartz, J., Liu, Y., van Donkelaar, A., Martin, R. V., Berrocal, V., & Bell, M. L. (2021). Examining PM2.5 concentrations and exposure using multiple models. *Environmental research*, *196*, 110432.

Kelp, M. M., Lin, S., Kutz, J. N., & Mickley, L. J. (2022). A new approach for determining optimal placement of PM2. 5 air quality sensors: case study for the contiguous United States. *Environmental Research Letters*, *17*(3), 034034.

Kerckhoffs, J., Hoek, G., Portengen, L. t., Brunekreef, B., & Vermeulen, R. C. (2019). Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental science & technology*, *53*(3), 1413-1421.

Kerckhoffs, J., Wang, M., Meliefste, K., Malmqvist, E., Fischer, P., Janssen, N. A., Beelen, R., & Hoek, G. (2015). A national fine spatial scale land-use regression model for ozone. *Environmental research*, *140*, 440-448.

Kim, S.-Y., Bechle, M., Hankey, S., Sheppard, L., Szpiro, A. A., & Marshall, J. D. (2020). Concentrations of criteria pollutants in the contiguous US, 1979–2015: Role of prediction model parsimony in integrated empirical geographic regression. *PloS one*, *15*(2), e0228535.

Knibbs, L. D., Coorey, C. P., Bechle, M. J., Cowie, C. T., Dirgawati, M., Heyworth, J. S., Marks, G. B., Marshall, J. D., Morawska, L., & Pereira, G. (2016). Independent validation of national satellite-based land-use regression models for nitrogen dioxide using passive samplers. *Environmental science & technology*, *50*(22), 12331-12338.

Knibbs, L. D., Van Donkelaar, A., Martin, R. V., Bechle, M. J., Brauer, M., Cohen, D. D., Cowie, C. T., Dirgawati, M., Guo, Y., & Hanigan, I. C. (2018). Satellite-based land-use regression for continental-scale long-term ambient PM2. 5 exposure assessment in Australia. *Environmental science & technology*, *52*(21), 12445-12455.

Lane, H. M., Morello-Frosch, R., Marshall, J. D., & Apte, J. S. (2022). Historical redlining is associated with present-day air pollution disparities in U.S. cities. *Environ Sci Technol Lett*, *9*(4), 345-350. https://doi.org/10.1021/acs.estlett.1c01012

Larkin, A., Anenberg, S., Goldberg, D. L., Mohegh, A., Brauer, M., & Hystad, P. (2023). A global spatial-temporal land use regression model for nitrogen dioxide air pollution. *Frontiers in Environmental Science*, *11*, 1125979.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., & Hystad, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environmental science & technology*, *51*(12), 6957-6964.

Lee, M., Brauer, M., Wong, P., Tang, R., Tsui, T. H., Choi, C., Cheng, W., Lai, P.-C., Tian, L., Thach, T.-Q., Allen, R., & Barratt, B. (2017). Land use regression modelling of air pollution in high density high rise cities: A case study in Hong Kong. *Science of the total environment*, *592*, 306-315. https://doi.org/https://doi.org/10.1016/j.scitotenv.2017.03.094

Li, L., Lurmann, F., Habre, R., Urman, R., Rappaport, E., Ritz, B., Chen, J.-C., Gilliland, F. D., & Wu, J. (2017). Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution. *Environmental science & technology*, *51*(17), 9920-9929.

Liu, J., Clark, L. P., Bechle, M. J., Hajat, A., Kim, S.-Y., Robinson, A. L., Sheppard, L., Szpiro, A. A., & Marshall, J. D. (2021). Disparities in air pollution exposure in the United States by race/ethnicity and income, 1990–2010. *Environmental health perspectives*, *129*(12), 127005. https://doi.org/doi:10.1289/EHP8584

Lloyd, M., Ganji, A., Xu, J., Venuta, A., Simon, L., Zhang, M., Saeedi, M., Yamanouchi, S., Apte, J., & Hong, K. (2023). Predicting spatial variations in annual average outdoor ultrafine particle concentrations in Montreal and Toronto, Canada: Integrating land use regression and deep learning models. *Environment International*, *178*, 108106.

Lotfata, A. (2022). Using geographically weighted models to explore obesity prevalence association with air temperature, socioeconomic factors, and unhealthy behavior in the USA. *Journal of Geovisualization and Spatial Analysis*, *6*(1), 14.

Lu, T., Bechle, M. J., Wan, Y., Presto, A. A., & Hankey, S. (2022). Using crowd-sourced low-cost sensors in a land use regression of PM2.5 in 6 US cities. *Air Quality, Atmosphere & Health*, *15*(4), 667-678.

Lu, T., Garcia, D. A., Garcia, A., & Liu, Y. (2023). Leveraging crowd-sourced environmental data to assess air pollution exposure disparity: A case of Los Angeles County. *International Journal of Applied Earth Observation and Geoinformation*, *125*, 103599.

Lu, T., Lansing, J., Zhang, W., Bechle, M. J., & Hankey, S. (2019). Land Use Regression models for 60 volatile organic compounds: Comparing Google Point of Interest (POI) and city permit data. *Science of the total environment*, *677*, 131-141.

Lu, T., Marshall, J. D., Zhang, W., Hystad, P., Kim, S.-Y., Bechle, M. J., Demuzere, M., & Hankey, S. (2021). National empirical models of air pollution using microscale measures of the urban environment. *Environmental science & technology*, *55*(22), 15519-15530.

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmospheric environment*, *42*(6), 1359-1369.

Masiol, M., Squizzato, S., Chalupa, D., Rich, D. Q., & Hopke, P. K. (2019). Spatial-temporal variations of summertime ozone concentrations across a metropolitan area using a network of low-cost monitors to develop 24 hourly land-use regression models. *Science of the total environment*, *654*, 1167-1178.

Masiol, M., Zíková, N., Chalupa, D. C., Rich, D. Q., Ferro, A. R., & Hopke, P. K. (2018). Hourly land-use regression models based on low-cost PM monitor data. *Environmental research*, *167*, 7-14. https://doi.org/https://doi.org/10.1016/j.envres.2018.06.052

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and intelligent laboratory systems*, *118*, 62-69.

Meng, X., Chen, L., Cai, J., Zou, B., Wu, C.-F., Fu, Q., Zhang, Y., Liu, Y., & Kan, H. (2015). A land use regression model for estimating the NO2 concentration in Shanghai, China. *Environmental research*, *137*, 308-315.

Messier, K. P., Chambliss, S. E., Gani, S., Alvarez, R., Brauer, M., Choi, J. J., Hamburg, S. P., Kerckhoffs, J., LaFranchi, B., & Lunden, M. M. (2018). Mapping air pollution with Google Street View cars: Efficient approaches with mobile monitoring and land use regression. *Environmental science & technology*, *52*(21), 12563-12572.

Nadal, M., Marquès, M., Mari, M., & Domingo, J. L. (2015). Climate change and environmental concentrations of POPs: A review. *Environmental research*, *143*, 177-185.

Nopmongcol, U., Beardsley, R., Kumar, N., Knipping, E., & Yarwood, G. (2019). Changes in United States deposition of nitrogen and sulfur compounds over five decades from 1970 to 2020. *Atmospheric environment*, *209*, 144-151. https://doi.org/https://doi.org/10.1016/j.atmosenv.2019.04.018

Novotny, E. V., Bechle, M., Millet, D. B., & Marshall, J. D. (2011). National satellite-based land use regression: NO2 in the United States. *Epidemiology*, *22*(1), S81.

Qi, M., Dixit, K., Marshall, J. D., Zhang, W., & Hankey, S. (2022). National land use regression model for no2 using street view imagery and satellite observations. *Environmental science & technology*, *56*(18), 13499-13509.

Qi, M., & Hankey, S. (2021). Using street view imagery to predict street-level particulate air pollution. *Environmental science & technology*, *55*(4), 2695-2704.

Roy, A. (2021). Atmospheric pollution retrieval using path radiance derived from remote sensing data. *Journal of Geovisualization and Spatial Analysis*, *5*(2), 26.

Saha, P. K., Hankey, S., Marshall, J. D., Robinson, A. L., & Presto, A. A. (2021). High-spatial-resolution estimates of ultrafine particle concentrations across the continental United States. *Environmental science & technology*, *55*(15), 10320-10331.

Saha, P. K., Presto, A. A., Hankey, S., Murphy, B. N., Allen, C., Zhang, W., Marshall, J. D., & Robinson, A. L. (2022). National exposure models for source-specific primary particulate matter concentrations using aerosol mass spectrometry data. *Environmental science & technology*, *56*(20), 14284-14295.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., & Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2. 5 concentrations in epidemiology. *Atmospheric environment*, *75*, 383-392.

Sanchez, M., Ambros, A., Milà, C., Salmon, M., Balakrishnan, K., Sambandam, S., Sreekanth, V., Marshall, J. D., & Tonne, C. (2018). Development of land-use regression models for fine particles and black carbon in peri-urban South India. *Science of the total environment*, *634*, 77-86.

Saraswat, A., Apte, J. S., Kandlikar, M., Brauer, M., Henderson, S. B., & Marshall, J. D. (2013). Spatiotemporal land use regression models of fine, ultrafine, and black carbon particulate matter in New Delhi, India. *Environmental science & technology*, *47*(22), 12903-12911.

Su, J. G., Brauer, M., Ainslie, B., Steyn, D., Larson, T., & Buzzelli, M. (2008). An innovative land use regression model incorporating meteorology for exposure analysis. *Science of the total environment*, *390*(2-3), 520-529.

US EPA. (2022). *Fused Air Quality Surface Using Downscaling (FAQSD)*. US EPA. Retrieved 2 October 2023 from https://www.epa.gov/hesc/rsig-related-downloadable-data-files

Van Donkelaar, A., Martin, R. V., Li, C., & Burnett, R. T. (2019). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental science & technology*, *53*(5), 2595-2611.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G., & Brunekreef, B. (2012). Systematic evaluation of land use regression models for NO2. *Environmental science & technology*, *46*(8), 4481-4489.

Wang, Y., Apte, J. S., Hill, J. D., Ivey, C. E., Johnson, D., Min, E., Morello-Frosch, R., Patterson, R., Robinson, A. L., & Tessum, C. W. (2023). Air quality policy should quantify effects on disparities. *Science*, *381*(6655), 272-274.

Weichenthal, S., Van Ryswyk, K., Goldstein, A., Bagg, S., Shekkarizfard, M., & Hatzopoulou, M. (2016). A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. *Environmental research*, *146*, 65-72.

Wong, P.-Y., Lee, H.-Y., Chen, Y.-C., Zeng, Y.-T., Chern, Y.-R., Chen, N.-T., Lung, S.-C. C., Su, H.-J., & Wu, C.-D. (2021). Using a land use regression model with machine learning to estimate ground level PM2. 5. *Environmental Pollution*, *277*, 116846.

Young, M. T., Bechle, M. J., Sampson, P. D., Szpiro, A. A., Marshall, J. D., Sheppard, L., & Kaufman, J. D. (2016). Satellite-based NO2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental science & technology*, *50*(7), 3686-3694.

Yuan, Z., Kerckhoffs, J., Hoek, G., & Vermeulen, R. (2022). A knowledge transfer approach to map long-term concentrations of hyperlocal air pollution from short-term mobile measurements. *Environmental science & technology*, *56*(19), 13820-13828. https://doi.org/10.1021/acs.est.2c05036

Zwack, L. M., Paciorek, C. J., Spengler, J. D., & Levy, J. I. (2011). Modeling spatial patterns of traffic-related air pollutants in complex urban terrain. *Environmental health perspectives*, *119*(6), 852-859.