 doi:10.20944/preprints202109.0164.v2

A High Resolution Spatiotemporal Fine Particulate Matter Exposure Assessment Model for the Contiguous United States

Cole Brokamp

Cincinnati Children's Hospital Medical Center; 3333 Burnet Ave, Cincinnati, OH, 45219

University of Cincinnati College of Medicine

Abstract

Currently available nationwide prediction models for fine particulate matter ($PM_{2.5}$) lack prediction confidence intervals and usually do not describe cross validated model performance at different spatiotemporal resolutions and extents. We used 41 different spatiotemporal predictors, including data on land use, meteorology, aerosol optical density, emissions, wildfires, population, traffic, and spatiotemporal indicators to train a machine learning model to predict daily averages of $PM_{2.5}$ concentrations at 0.75 sq km resolution across the contiguous United States from 2000 through 2020. We utilized a generalized random forest model that allowed us to generate asymptotically-valid prediction confidence intervals and took advantage of its usefulness as an ensemble learner to quickly and cheaply characterize leave-one-location-out CV model performance for different temporal resolutions and geographic regions. Using a variable importance metric, we selected 8 predictors that were able to accurately predict daily $PM_{2.5}$, with an overall leave-one-location-out cross validated median absolute error of 1.20 $\frac{\mu g}{m^3}$, an R^2 of 0.84, and confidence interval coverage fraction of 95%. When considering aggregated temporal windows, the model achieved leave-one-location-out cross validated median absolute errors of 0.99, 0.76, 0.63, and 0.60 $\frac{\mu g}{m^3}$ for weekly, monthly, annual, and all-time exposure assessments, respectively. We further describe the model's cross validated performance at different geographic regions in the United States, finding that it performs worse in the Western half of the country where there are less monitors. The code and data used to create this model are publicly available and we have developed software packages to be used for exposure assessment. This accurate exposure assessment model will be useful for epidemiologists seeking to study the health effects of $PM_{2.5}$ across the continental United States, while possibly considering exposure estimation accuracy and uncertainty specific to the the spatiotemporal resolution and extent of their study design and population.

Keywords

Fine particulate matter; exposure assessment; machine learning; spatiotemporal; high resolution

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

Introduction

Particulate matter less than 2.5 μm ($PM_{2.5}$) has a wide range of negative health effects, including on the respiratory system, cardiovascular system, metabolic system, nervous system, reproductive system, pregnancy and birth outcomes, cancer, and mortality ("Integrated Science Assessment for Particulate Matter" 2020). Newer developments in $PM_{2.5}$ exposure assessment methods have improved the spatial resolution and accuracy of estimates, often reducing the bias and uncertainty present in health studies. Furthermore, the relationship between $PM_{2.5}$ and health effects has been studied at increasingly higher temporal resolutions, allowing for investigation of its short-term and acute effects. Application of these models within large registries and multi-site studies has highlighted the need for timely, accessible, open, high resolution, and nationwide exposure assessment models that can quickly be tailored to individual studies at different spatiotemporal resolutions and extents. Current nationwide $PM_{2.5}$ exposure assessment models do not consider model performance at different spatiotemporal aggregations, resolutions, and extents (e.g., daily estimates at zip codes across the entire United States versus annual averages at exact locations within one city), which would allow for study-specific estimates of bias and uncertainty. Furthermore, uncertainty in exposure assessment of ambient pollutants is often ignored in health studies because exposure predictions do not include prediction standard errors or prediction 95% confidence intervals. Without this critical feature, scientists cannot propogate uncertainty from exposure assessment to health effects models in order to produce unbiased estimates that truly reflect the uncertainty contributed by both exposure estimation and health effects modeling.

The objective here was to create an open and reusable high resolution (0.75 sq km) spatiotemporal $PM_{2.5}$ exposure assessment model for the contiguous United States from 2000 through 2020. We used existing $PM_{2.5}$ measurements from the Environmental Protection Agency to train a prediction model based on features related to land use, meteorology, aerosol optical density, emissions, wildfires, population density, development, and traffic. We designed this model such that it could (1) produce accurate predictions with accompanying estimates of prediction uncertainty and (2) be used to easily quantify cross validated accuracy at different spatial and temporal resolutions and extents.

Methods

Study Domain and Spatiotemporal Grid

In order to spatially harmonize predictors and predictions, the H3 hexagonal hierarchical spatial index (Brodsky 2018) at a resolution of 8 was used to create a grid of 11,932,970 hexagonal cells across the contiguous United States. Each hexagon grid cell is represented using a geohash of 15 characters and has an average area of 0.74 sq km and an average side length of 461.4 m. A hierarchical spatial index was used to quickly scale spatial data and because an average polygon (e.g., zip code or census tract) can be approximated with hexagonal tiles with a smaller margin of error than would be possible with square tiles (Brodsky 2018).

All spatiotemporal data was collected when available from 2000 through 2020 and was harmonized as a daily average. Spatial calculations to derive features (e.g., distance, area) were completed in the Conus Albers NAD83(NSRS2007) projection.

Data

$PM_{2.5}$ Measurements

Observed 24-hour average $PM_{2.5}$ measurements for the contiguous United States from 2000 through 2020 were retrieved from the Environmental Protection Agency (EPA) Air Quality System (AQS). Co-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

located daily measurements were averaged. The latitude and longitude for each measurement location was geohashed to its containing resolution 8 H3 grid cell.

Land Use

Land use information was obtained from the National Land Cover Database (NLCD) and summarized at each resolution 8 H3 grid cell as the average percentage imperviousness of containing 30 sq m cells. We also included the fraction of each cell that was "green" (i.e. any NLCD category except water, ice/snow, developed medium intensity, developed high intensity, rock/sand/clay) as well as the fraction of each cell that was non-impervious. The impervious category was further classified into types of imperviousness, including primary urban, primary rural, secondary urban, secondary rural, tertiary urban, tertiary rural, thinned urban, thinned rural, nonroad urban, nonroad rural, energy production urban, energy production rural. NLCD data was available as an annual value, available for 2001, 2006, 2011, and 2016. In the training and prediction data, each day was assigned to the nearest calendar year with available data.

Meteorology

Meteorological information was obtained from the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) dataset (Kistler et al. 2001). Daily averages of planetary boundary layer height, visibility, wind speed and direction, air temperature, relative humidity, precipitation rate, and surface pressure for each resolution 8 H3 grid cell were calculated based on the 12 x 12 km NARR cell that contained its centroid.

Aerosol Optical Density

Aerosol optical density (AOD) information was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard the National Aeronautics and Space Administration (NASA) Terra and Aqua satellites. Specifically, AOD at 470 nm from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) Land AOD Daily L2G 1km SIN Grid V006 product (MCD19A2) (https://doi.org/10.5067/MODIS/MCD19A2.006) was used by joining the centroid of each 1 km grid cell to its containing resolution 8 H3 hexagon cell. AOD measurements were utilized if they were (1) less than 2, (2) were from pixels that were "Clear" or "Possibly Cloudy" in the Cloud Mask, and (3) were from pixels that were "Clear" or "Adjacent to a single cloudy pixel" in the Adjacency Mask.

$PM_{2.5}$ Emissions

Emissions information was obtained from the National Emissions Inventory (NEI) Database as annual estimated tons of $PM_{2.5}$, available for 2008, 2011, 2014, and 2017. County-level emissions due to nonpoint, onroad, nonroad, and event sources were joined to all containing resolution 8 H3 hexagon cells. NEI event sources capture wildfires and prescribed burns and are calculated by the EPA based on satellite detection approaches, fire models, and fire activity data provided by local agencies. The exact latitude and longitude of point sources were assigned to their containing resolution 8 H3 hexagon cells and all cells were assigned a value equal to the distance to the closest NEI site. In the training and prediction data, each day was assigned to the nearest calendar year with data.

Wildfires

Wildfire information was obtained from Version 1.5 of the Fire Inventory from the National Center for Atmospheric Research (FINN) dataset (Wiedinmyer et al. 2011) that estimates daily open biomass burning at a 1 km sq resolution as the area of total land burned and the estimated $PM_{2.5}$ emitted from the fire event. The centroid of each 1 km grid cell was joined to its containing resolution 8 H3 hexagon cell, with daily estimates available for 2002 through 2018.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

Population Density

Estimates of total population from the 2018 5-yr American Community Survey at the census tract level were used to calculate the population density for each resolution 5 H3 hexagon cell as the area-weighted total population divided by the total area.

Traffic

Distance to the closest S1100 road from the 2018 TIGER/Line Shapefiles for the centroid of each resolution 8 H3 hexagon cell centroid was used to approximate interstate traffic.

Derived Spatiotemporal Predictors

To include temporal information, we included the calendar year, day of the year, and day of the week as numeric predictors as well as an indicator variable that was true if any date was a U.S. holiday (New Year's Day, Fourth of July, Thanksgiving, Christmas, Labor Day, Memorial Day, Martin Luther King Jr. Day). Spatial predictors were included as the X and Y coordinates of each grid cell and an identifier for each H3 grid cell.

A convolution function was used to create a training feature describing average regional daily $PM_{2.5}$ by spatially aggregating the measurements to resolution 3 H3 grid cells (average area of 12,392 sq km each), calculating the inverse distance weighted mean of neighboring cells within a 5-ring buffer, and then calculating a moving three day average.

Generalized Random Forest Modeling

We used the generalized random forest (GRF) framework (Athey, Tibshirani, and Wager 2019) to train a regression forest. This differs from the more traditional random forest by (1) sampling without replacement for each resample in order to be able to generate variance estimates and (2) using the gradient of the objective function to optimize a linear approximation to the criterion in order to reduce computation times. Within the GRF, variances for predictions are estimated by training trees in small groups on subsets of each resample and then comparing predictions within and across groups (Athey, Tibshirani, and Wager 2019).

We considered autocorrelation among measurements from the same locations in order to produce cluster-robust predictions and variance estimates by resampling at the location- rather than the measurement-level (i.e., location-day-level). This was accomplished by using the H3 grid cell identifier to select observations for resamples such that all observations within one location were selected during training and predictions.

Missing data in predictors was handled using the missing incorporated in attributes criterion (Twala, Jones, and Hand 2008) because our previous work has shown that missingness itself, especially within AOD measurements, can be a good predictor of $PM_{2.5}$ concentrations (Brokamp, Jandarov, et al. 2018). In practice, this means that observations with missing values can be split on missingness and/or their observed values.

Random forests were trained using a subsample of half the available training data for each tree and each split was generated by considering 14 randomly selected predictors, with the restrictions that each resulting split could not be less than 5% of the observations in the parent node and must have at least 1 observation. We trained one random forest model per calendar year for the entire contiguous United States.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

Assessing Cross Validated Model Performance

Leave one location out (LOLO) cross validated (CV) predictions were created for each measured daily $PM_{2.5}$ concentration by holding out one location and predicting all of its daily measurements using the remaining locations. To estimate model performance when averaging concentrations over weeks, months, years, or all time, we averaged the measured concentrations and predictions and defined the standard error of the average means as the square root of the sum of the squared individual standard errors. Cross validated statistics for each of the held out locations and temporal windows (daily, weekly, monthly, annual, all) were calculated, including median absolute error (MAE), root mean squared error (RMSE), squared correlation between predicted and observed (R^2), slope from an intercept-less regression model between predicted and observed (slope), and the percentage of measurements that were covered by the 95% prediction confidence interval. The cross validated statistics across all locations for each temporal window were summarized using the median and boxplots.

Variable Selection

We initially used a variable importance metric (a weighted sum of how many times each feature was split on at each depth in the forest, up to 4 splits deep) to sequentially reduce the number of predictors included in the GRF model while estimating the cross validated MAE. At each step, we removed variables that were at least one order of magnitude lower in importance compared to the most important variable. We balanced the loss in accuracy against the computational time needed to train and predict using the GRF model as well as the interpretibility of the included predictors.

Computing

All aspects of this analysis were completed in R (R Core Team 2021), specifically using the sf (Pebesma 2018), h3 (Kuethe 2019), fst (Klik 2020) and grf (Tibshirani, Athey, and Wager 2020) packages. All code used to acquire and harmonize data, train and evaluate models, and make predictions is publicly available in an online repository (https://github.com/geomarker-io/st_pm_hex).

Results and Discussion

Training Data

Data on measured $PM_{2.5}$ concentrations and 41 different predictors, including land use, meteorology, AOD, $PM_{2.5}$ emissions, wildfires, population density, traffic, and spatiotemporal indicators were harmonized to daily averages within cells in a hexagonal lattice pattern defined by the H3 hexagonal hierarchical spatial index (Brodsky 2018). This resulted in 11,932,970 hexagonal cells covering the contiguous United States, with an average area each of 0.74 sq km. The hierarchical nature of the index system allowed us quickly join indexed data across disparate spatial resolutions and aggregate them across different levels of precision as needed. The hexagon is preferable to other polygons that tile regularly (i.e., square and triangle) because its immediate neighbors are equidistant and expanding rings of neighbors approximate circles, which reduces spatial distortion and misalignment problems when scaling spatial data in a hierarchical grid or lattice.

2,374,589 24-hour average $PM_{2.5}$ measurements from 2000 through 2020 contained within 1,685 unique hexagon grid cells were retrieved. Grid cells with $PM_{2.5}$ measurements had a median of 1,042 distinct daily measurements and each calendar day with $PM_{2.5}$ measurements had a median of 150 grid cells. Daily measurements ranged from -0.1 to 640.6 $\frac{\mu g}{m^3}$, with a mean of 10.76 $\frac{\mu g}{m^3}$ (25th quartile: 5.80, median: 9.00, 75th quartile: 13.60 $\frac{\mu g}{m^3}$). Notably, only 0.3% ($n = 7{,}184$) of grid-days with $PM_{2.5}$ measurements had

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

non-missing AOD measurements. Table 3 contains the number of $PM_{2.5}$ measurements and the mean $PM_{2.5}$ concentration for each calendar year.

Generalized Random Forest Modeling

We trained regression forests using the generalized random forest (GRF) framework to predict measured $PM_{2.5}$ concentrations. Variable importance was used to select the following predictors (listed in order of decreasing importance in the final model): $PM_{2.5}$ convolution, X coordinate, day of year, air temperature, $PM_{2.5}$ emissions event, planetary boundary layer height, Y coordinate, and wind speed in the v direction. The variable importance for the initial regression forest containing all predictors in presented in Table 1. The cross validated MAE for this forest was 0.97 $\frac{\mu g}{m^3}$. Regression forests trained with 28, 14, and 8 predictors with the highest variable importances had cross validated MAEs of 0.97, 0.93, and 0.90 $\frac{\mu g}{m^3}$, respectively.

Table 1*: Variable importance for the initial regression forest trained on $PM_{2.5}$ concentrations using 41 predictors.*

variable importance	variable
0.345	$PM_{2.5}$ convolution
0.178	X coordinate
0.111	planetary boundary height
0.083	day of year
0.054	$PM_{2.5}$ event emissions
0.048	wind speed in v direction
0.043	air temperature
0.036	relative humidity
0.018	Y coordinate
0.016	visibility
0.013	$PM_{2.5}$ nonroad emissions
0.012	wind speed in u direction
0.008	pressure
0.008	population density
0.006	distance to nearest $PM_{2.5}$ emission point source
0.006	precipitation rate
0.004	fraction impervious land
0.003	fraction nonimpervious land
0.002	$PM_{2.5}$ point emissions, fraction nonroad urban land, distance to nearest major roadway, fraction green land
0.001	holiday, day of week, fraction tertiary urban land, aerosol optical density
0.0000	fraction of land category (primary urban, primary rural, secondary urban, secondary rural, tertiary rural, thinned urban land, thinned rural, nonroad rural, energy production urban, energy production rural), year, $PM_{2.5}$ onroad emissions, $PM_{2.5}$ nonpoint emissions, fire $PM_{2.5}$ emissions, fire area

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

Our variable importance results suggest that aerosol optical density did not significantly increase $PM_{2.5}$ prediction accuracy, similar to a previous nationwide $PM_{2.5}$ prediction model (Di et al. 2019) and a research report on this question (Paciorek and Liu 2012). Notably, the initial model with 41 predictors had an out of bag MAE of 0.97 $\frac{\mu g}{m^3}$ and was reduced to 8 total predictors while maintaining a similarly low out of bag MAE of 0.90 $\frac{\mu g}{m^3}$. Utilizing a lower number of predictors in the final model reduces the computational resources required to train and predict with machine learning models. Furthermore, although decreases in exposure prediction error do not necessarily lead to less biased health effects estimates, (Szpiro, Paciorek, and Sheppard 2011) including exposure predictors as covariates in health effects models can cause bias (Cefalu and Dominici 2014). However, there are methods proposed for correcting inference based on outcomes predicted by machine learning (Wang, McCormick, and Leek 2020) and specifically within air pollution epidemiology (Szpiro, Sheppard, and Lumley 2011; Szpiro and Paciorek 2013; Gryparis et al. 2008).

Quantifying Model Performance

Although leave-one-location-out (LOLO) CV estimates have lower bias and variance compared to k-fold approaches (Watson et al. 2020), the latter are often used due to computationally-intensive repetitive model training on data subsets or resamples. Here, we were able to estimate LOLO CV predictions without requiring expensive iterative model training by making predictions using only trees from the generalized random forest that did not use the held-out location for training.

This allowed for the use of a single trained ensemble to generate LOLO estimates for 1,685 unique locations and proved to be an advantage over other predictive machine learning models that do not utilize this ensemble framework of bagged aggregation, such as boosted trees and neural networks. Instead of using one error statistic to summarize model performance, we used this novel method to quickly and cheaply characterize LOLO prediction error for different temporal resolution (e.g., daily, weekly, monthly, annual) and geographic regions.

The median of the cross validated statistics across all locations for each temporal window are presented in Table 2 and the distribution of MAE and R^2 are illustrated as boxplots in Figure 1. The cross validation analysis suggested that the model made highly accurate predictions, with a daily R^2 of 0.84, MAE of 1.20 $\frac{\mu g}{m^3}$, and a 95% confidence interval coverage of 95%. When averaging cross validated predictions and measurements across time, the model R^2 increased for weekly (0.86), monthly (0.90), and annual (0.95) temporal windows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

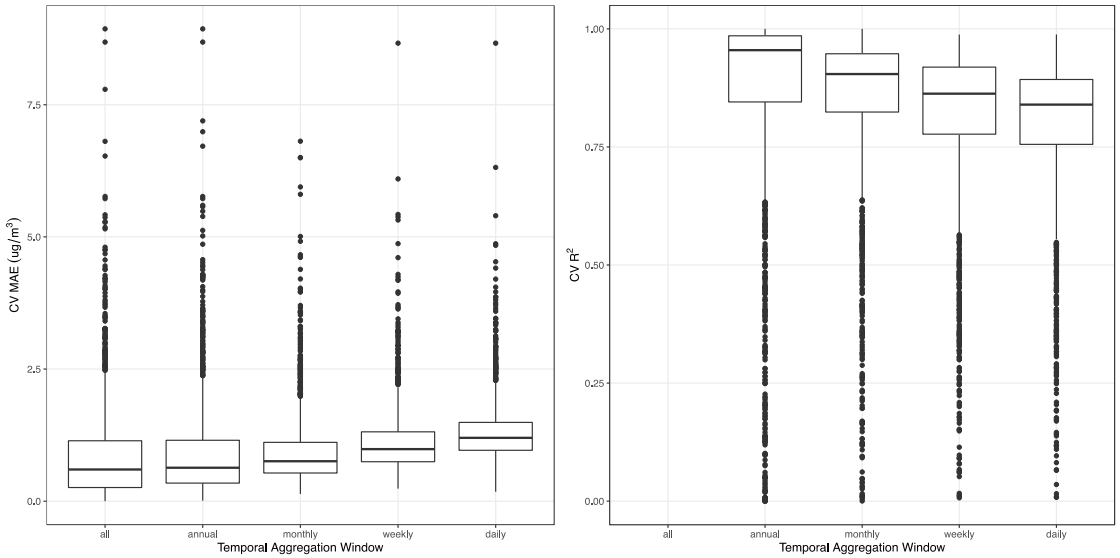

***Figure 1**: Boxplot of cross validated estimates of R^2 and MAE across 1,685 locations where $PM_{2.5}$ was measured for different temporally aggregated predictions and measurements.*

***Table 2**: Median leave one location out (LLO) cross validated (CV) error estimates (MAE: median absolute error, RMSE: root mean squared error) for 1,685 spatial cells with measured average $PM_{2.5}$ concentrations for different temporal windows. N is the median number of temporal windows per location used to calculate the median CV error statistics.*

Temporal Average	*N*	MAE	RMSE	R^2	Slope	95% CI Coverage
all	1	0.60	0.60	-	1.01	100%
annual	10	0.63	0.78	0.95	1.00	100%
monthly	110	0.76	1.28	0.90	0.99	100%
weekly	466	0.99	2.03	0.86	0.98	99%
daily	1,042	1.20	2.61	0.84	0.96	95%

Because model performance was expected to change over time, LOLO CV accuracy was estimated for daily predictions by year from 2000 through 2020 (Table 3). Even though the number of overall measurements decreased over time, the overall concentration and variability of $PM_{2.5}$ also decreased over time, resulting in drastic changes in model improvement for any one given year (Table 3).

***Table 3**: Mean and number of daily $PM_{2.5}$ measurements used to train the predictive model along with annual leave one location out (LLO) cross validated (CV) error estimates (MAE: median absolute error) for each year.*

Year	N	Mean $PM_{2.5}$	MAE	R^2
2000	142,124	13.40	1.45	0.81
2001	153,656	12.97	1.37	0.80
2002	156,093	12.52	1.28	0.84
2003	140,008	12.38	1.33	0.82
2004	137,625	12.06	1.28	0.82
2005	133,048	12.89	1.35	0.84
2006	125,636	11.72	1.25	0.82
2007	134,178	12.17	1.28	0.83

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

2008	128,651	10.98	1.17	0.80
2009	128,288	9.91	1.06	0.79
2010	120,322	10.05	1.07	0.81
2011	99,527	10.03	1.13	0.80
2012	97,985	9.29	1.09	0.78
2013	98,106	9.06	0.99	0.79
2014	97,810	8.95	1.03	0.79
2015	102,986	8.61	0.97	0.81
2016	99,278	7.98	0.90	0.77
2017	90,690	8.26	0.90	0.79
2018	82,222	8.43	0.93	0.78
2019	74,660	7.82	0.92	0.78
2020	72,055	7.61	0.86	0.85

Because model performance was expected to vary by location, LOLO CV MAE was estimated for each time window separately for 81 different regions corresponding to resolution 2 H3 cells covering the study domain (average area for each cell of 86.7 sq km; Figure 2). The model tended to perform better in the Eastern Half of the United States, which had more $PM_{2.5}$ measurements. Notably 3 of the regions in the Western United States did not have any $PM_{2.5}$ measurements and so cross validated error estimates could not be calculated for these locations. Estimation of model accuracy for predicted exposures in different regions of the United States as well as at different temporal aggregations will allow health studies to consider the impact of model accuracy specific to their spatiotemporal domain; for example, daily exposures within one major metropolitain area versus monthly exposures across the entire United States.

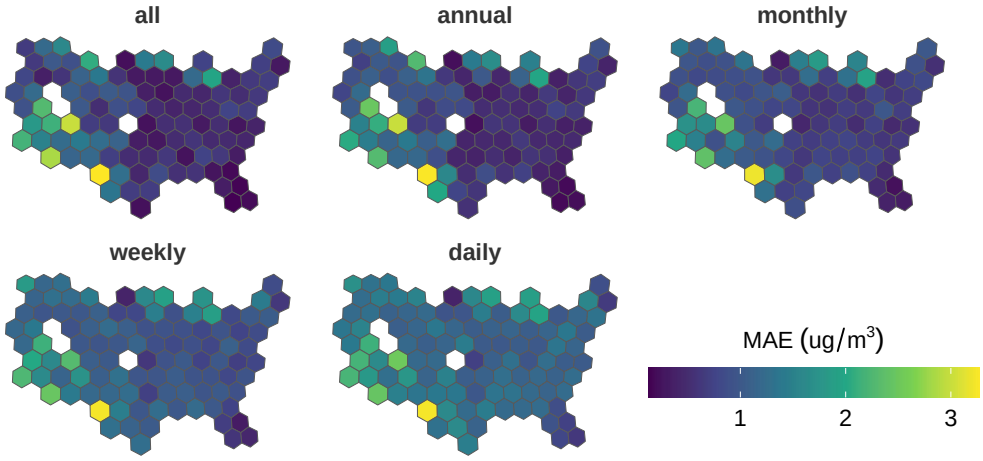

***Figure 2**: Leave one location out (LOLO) cross validated (CV) median absolute error (MAE) estimated for each time window separately for 81 different regions of the United States corresponding to resolution 2 H3 cells; the average area for each cell is 86.7 sq km.*

$PM_{2.5}$ Predictions

Figure 3 shows the relationship between all cross validated model predictions and $PM_{2.5}$ observations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

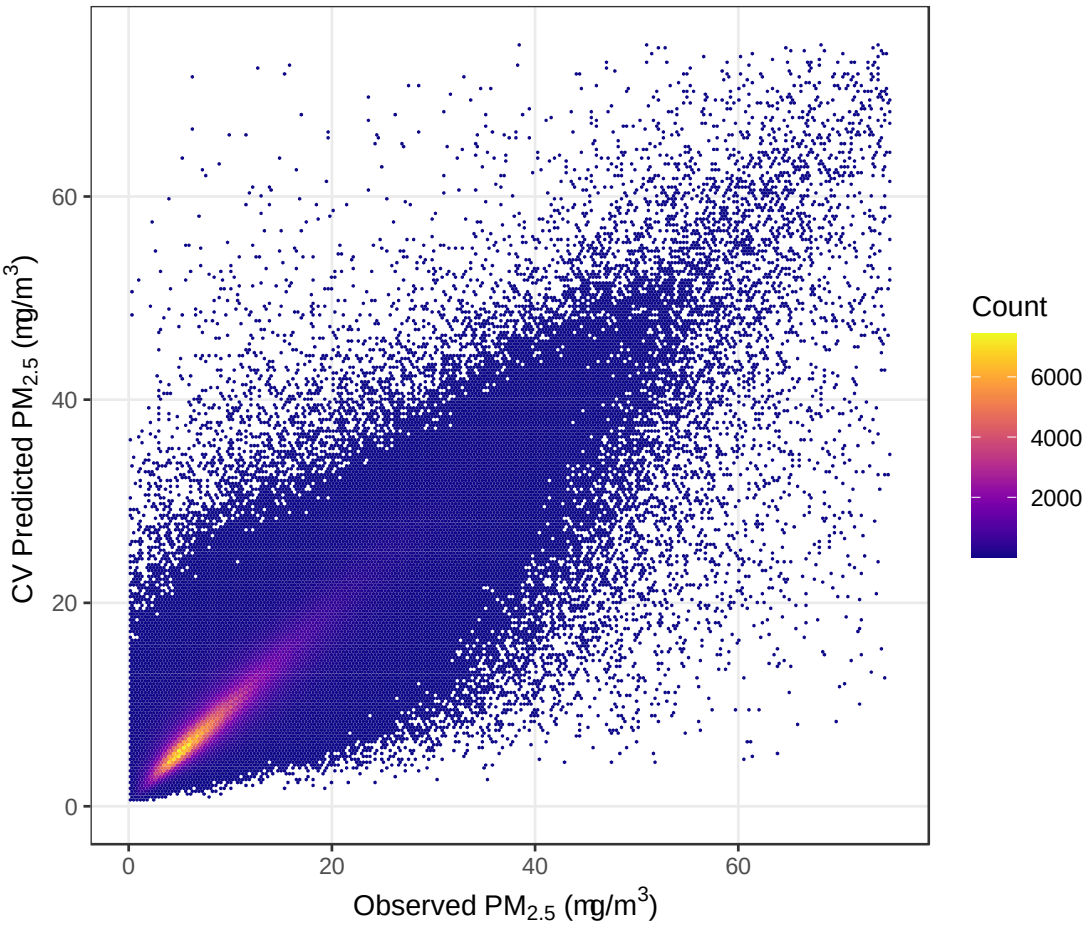

***Figure 3**: A 2-D histogram of all cross validated model predictions and $PM_{2.5}$ observations. Each hexagon bin is colored according to the number of points contained within it.*

Figure 4 presents the relationship between the length of the predicted $PM_{2.5}$ confidence interval and the residual for 1,000 randomly selected cross validated predictions. As expected, the model is more confident (i.e., shorter prediction 95% confidence intervals) when it is more accurate (i.e., smaller residual values).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

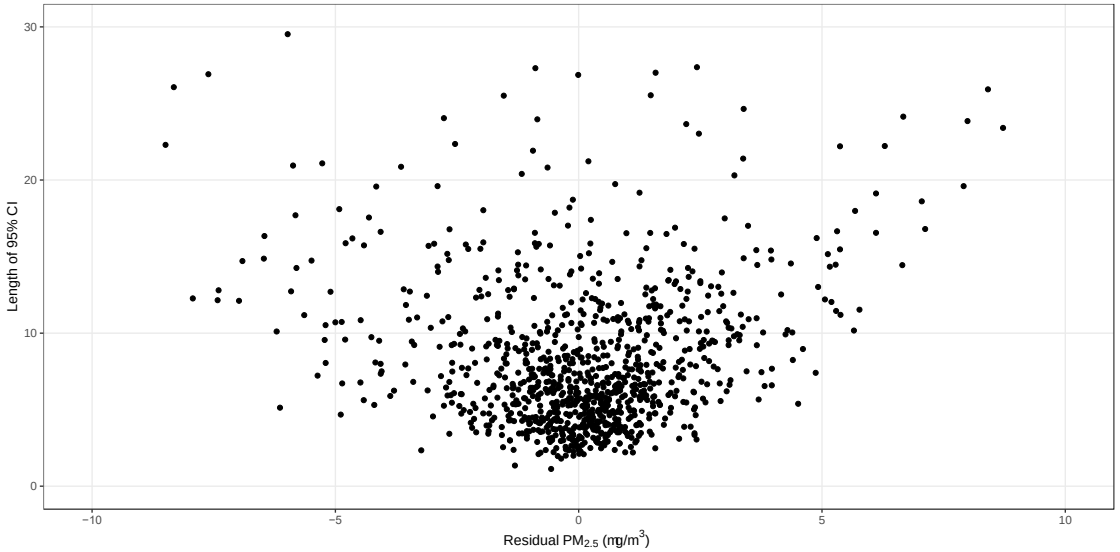

***Figure 4**: The relationship between the length of the predicted $PM_{2.5}$ confidence interval and the residual for 1,000 randomly selected cross validated predictions.*

The final GRF model was used to predict daily $PM_{2.5}$ estimates and their accompanying standard errors from 2000 through 2020 at each of the resolution 8 H3 cells covering the contiguous United States. These predictions are currently available publicly online in several files organized by geographic region and calendar year. Software packages that scientists can use for secure, private, and reproducible $PM_{2.5}$ assessment at scale are available, including an R package (https://geomarker.io/addPmData) and a DeGAUSS (Brokamp, Wolfe, et al. 2018) container (https://degauss.org/pm).

To demonstrate the spatiotemporal precision and variability in predicted $PM_{2.5}$ concentrations, we chose three urban locations (Boston MA, Cincinnati OH, Los Angeles CA) and created maps for three days from different seasons in 2019 (Figure 5) as well as time series plots for all days in 2019 (Figure 6). Predicted $PM_{2.5}$ concentrations varied greatly day-to-day between and within the three example cities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

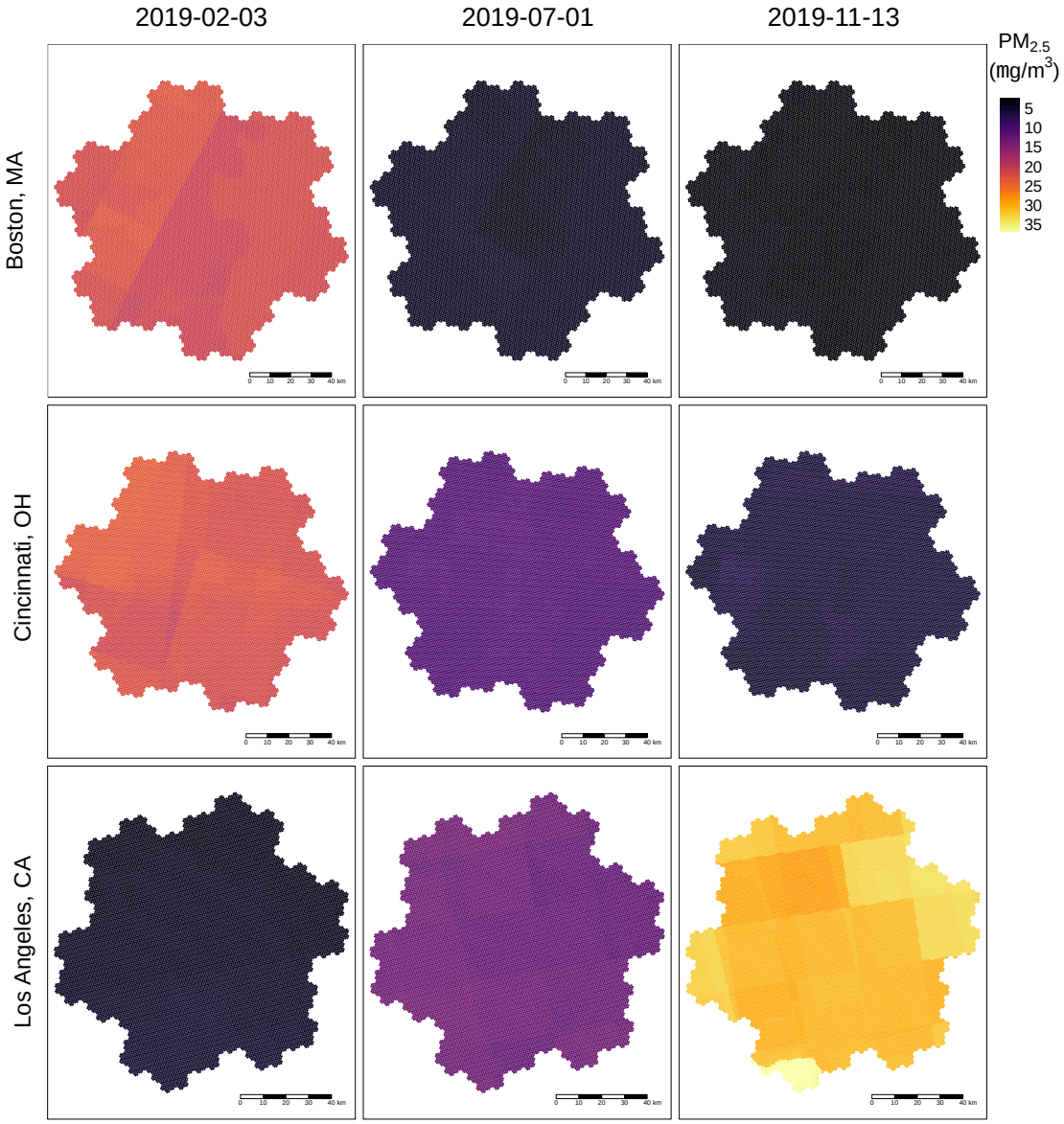

***Figure 5**: Predicted $PM_{2.5}$ concentrations at three urban locations for three days from different seasons in 2019. Each map panel depicts all resolution 8 H3 cells that makeup the one resolution 3 H3 cell covering the respective locations. The location of each city within the United States is shown in an inset map in Figure 6.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2021 doi:10.20944/preprints202109.0164.v2

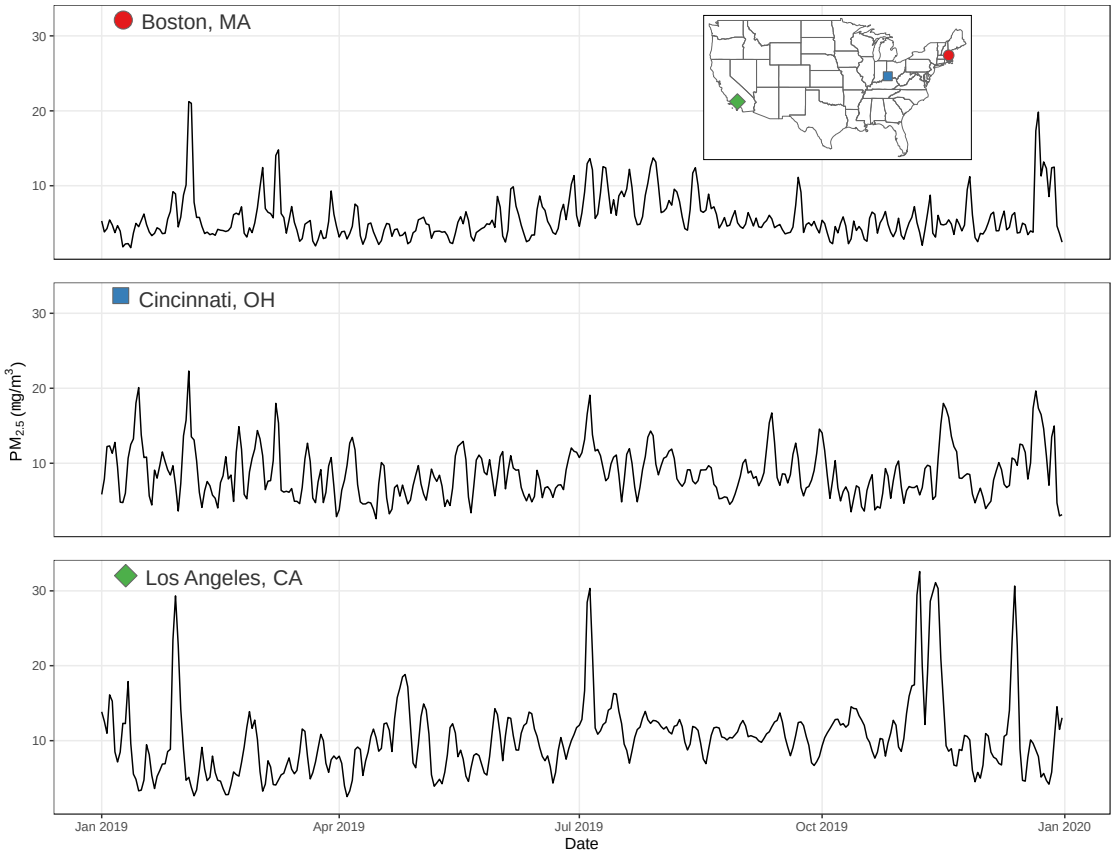

***Figure 6**: Daily predicted $PM_{2.5}$ concentrations at three urban locations for 2019. The inset map depicts the location of each city within the United States.*

Discussion

This work is the first $PM_{2.5}$ exposure assessment model with prediction confidence intervals and will allow for use of existing approaches (Szpiro and Paciorek 2013; Spiegelman 2016; Zandbergen et al. 2012) to propogate uncertainty from exposure estimation to health effect estimation. Other approaches have taken an empirical approach to estimating prediction uncertainty by using the standard deviation of predictions from multiple models or of prediction residuals, but this model is the first to produce asymptotically-valid prediction confidence intervals. Additionally, our novel approach to LOLO CV will allow for quick estimation of model accuracy at different spatial and temporal resolutions and extents that can be specialized to specific study populations.

Compared to other recent nationwide $PM_{2.5}$ prediction models, our model demonstrated a similar level of cross validated accuracy, with an LOLO daily CV R^2 of 0.84 compared to 10-fold CV R^2 estimates of 0.86 (Di et al. 2019) and 0.84 (Hu et al. 2017). Notably, our cross validated predictions at the annual level showed an improved performance (LOLO CV R^2: 0.95) compared to another existing model that also assessed annual performance (Di et al. 2019) (10-fold CV R^2: 0.89). Small differences in model performances are to be expected given that our model included more recent predictions (i.e. 2018 - 2020) and that we used a leave-one-location-out rather than leave-10-fold-out approach to estimate cross validated model performance.

One limitation inherent to our approach is possible spatiotemporal discontinuities in predicted concentrations due to the use of spatiotemporal data with different resolutions and frequencies. One example of this can be seen in Figure 5, specifically predictions for November 13th, 2019 in Los Angeles, CA show discontinuities coinciding with the boundaries of the NARR meteorological data. Predictions

 doi:10.20944/preprints202109.0164.v2

could be linearly (or non-linearly) smoothed if the small introduction of bias would be outweighed by removing discontinuities.

In the future, the use of the hexagonal hierarchical spatial indexing system will be advantageous for querying model predictions at different spatial aggregation scales, which might be necessary for protection of private health information in epidemiologic studies. Furthermore, the hexagonal hierarchical spatial index was designed to minimize quantization error introduced when people move throughout cities, making this exposure assessment tool ideal for utilizing high resolution GPS or travel activity data in order to assess high resolution daily ambient $PM_{2.5}$ exposures for health effects studies. The increasingly higher temporal precision at which health outcomes are measured will make temporal environmental health studies better, but will require hourly resolution exposure assessment models, as will be feasible with the upcoming NASA TEMPO satellite (Zoogman et al. 2017) and that has been done elsewhere (Jiang et al. 2021; Sun, Gong, and Zhou 2021). Future research could work to extend this modeling framework to hourly $PM_{2.5}$ concentrations by utilizing TEMPO satellite data and hourly meteorological data. Additionally the current model framework could also be expanded to other EPA criteria pollutants.

As this set of predictions and prediction model is maintained and used for exposure assessment and methodology, we aim to make it findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al. 2016). In addition to making the code and data used to create the model publicly available, we are also making estimates freely available and have developed software packages to be used for exposure assessment. This free and open source exposure prediction model could provide more insight and transparency into human health studies that are eventually used for policy decisions, including the EPA's National Ambient Air Quality Standards. FAIR model predictions and software tools will hopefully make high resolution exposure assessment for human health studies more easily and widely used. Ideally, this would reduce barriers of entry for scientists interested in environmental exposures and would facilitate more interdisciplinary research in existing and ongoing nationwide health studies.

Conclusion

In conclusion, this accurate exposure assessment model will be useful for epidemiologists seeking to study the health effects of $PM_{2.5}$ across the continential United States, while possibly considering exposure estimation accuracy and uncertainty specific to the spatiotemporal resolution and extent of their study design and population.

Acknowledgements

This work was supported by the National Institutes of Health (award numbers R01LM013222 and R01ES031621). The author declares no competing financial interest. Thanks to Dr. Patrick Ryan and Ms. Erika Rasnick for feedback on early versions of this work.

References

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–78.

Brodsky, Isaac. 2018. "H3: Hexagonal Hierarchical Geospatial Indexing System." *Uber Open Source. Retrieved*.

Brokamp, Cole, Roman Jandarov, Monir Hossain, and Patrick Ryan. 2018. "Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model." *Environmental Science & Technology* 52 (7): 4173–79. https://doi.org/10.1021/acs.est.7b05381.

Brokamp, Cole, Chris Wolfe, Todd Lingren, John Harley, and Patrick Ryan. 2018. "Decentralized and Reproducible Geocoding and Characterization of Community and Environmental Exposures for Multisite Studies." *Journal of the American Medical Informatics Association* 25 (3): 309–14.

Cefalu, Matthew, and Francesca Dominici. 2014. "Does Exposure Prediction Bias Health Effect Estimation? The Relationship Between Confounding Adjustment and Exposure Prediction." *Epidemiology (Cambridge, Mass.)* 25 (4): 583.

Di, Qian, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M Benjamin Sabath, et al. 2019. "An Ensemble-Based Model of Pm2. 5 Concentration Across the Contiguous United States with High Spatiotemporal Resolution." *Environment International* 130: 104909.

Gryparis, A., C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull. 2008. "Measurement Error Caused by Spatial Misalignment in Environmental Epidemiology." *Biostatistics* 10 (2): 258–74. https://doi.org/10.1093/biostatistics/kxn033.

Hu, Xuefei, Jessica H Belle, Xia Meng, Avani Wildani, Lance A Waller, Matthew J Strickland, and Yang Liu. 2017. "Estimating Pm2. 5 Concentrations in the Conterminous United States Using the Random Forest Approach." *Environmental Science & Technology* 51 (12): 6936–44.

"Integrated Science Assessment for Particulate Matter." 2020. *Fed. Regist.* https://www.federalregister.gov/documents/2020/01/27/2020-01223/integrated-science-assessment-for-particulate-matter.

Jiang, Tingting, Bin Chen, Zhen Nie, Zhehao Ren, Bing Xu, and Shihao Tang. 2021. "Estimation of Hourly Full-Coverage Pm2.5 Concentrations at 1-Km Resolution in China Using a Two-Stage Random Forest Model." *Atmospheric Research* 248 (January): 105146. https://doi.org/10.1016/j.atmosres.2020.105146.

Kistler, Robert, Eugenia Kalnay, William Collins, Suranjana Saha, Glenn White, John Woollen, Muthuvel Chelliah, et al. 2001. "The NCEP NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation." *Bulletin of the American Meteorological Society* 82 (2): 247–68.

Klik, Mark. 2020. *Fst: Lightning Fast Serialization of Data Frames for r*. https://CRAN.R-project.org/package=fst.

Kuethe, Stefan. 2019. *H3: R Bindings for H3*. https://github.com/crazycapivara/h3-r.

Paciorek, Christopher J, and Yang Liu. 2012. "Assessment and Statistical Modeling of the Relationship Between Remotely Sensed Aerosol Optical Depth and Pm2. 5 in the Eastern United States." *Research Report (Health Effects Institute)*, no. 167: 5–83.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spiegelman, Donna. 2016. "Evaluating Public Health Interventions: 4. The Nurses' Health Study and Methods for Eliminating Bias Attributable to Measurement Error and Misclassification." *American Journal of Public Health* 106 (9): 1563–66.

Sun, Jin, Jianhua Gong, and Jieping Zhou. 2021. “Estimating Hourly Pm2. 5 Concentrations in Beijing with Satellite Aerosol Optical Depth and a Random Forest Approach.” *Science of The Total Environment* 762: 144502.

Szpiro, Adam A., and Christopher J. Paciorek. 2013. “Measurement Error in Two-Stage Analyses, with Application to Air Pollution Epidemiology.” *Environmetrics* 24 (8): 501–17. https://doi.org/10.1002/env.2233.

Szpiro, Adam A., Christopher J. Paciorek, and Lianne Sheppard. 2011. “Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates?” *Epidemiology* 22 (5): 680–85. https://doi.org/10.1097/ede.0b013e3182254cc6.

Szpiro, Adam A., Lianne Sheppard, and Thomas Lumley. 2011. “Efficient Measurement Error Correction with Spatially Misaligned Data.” *Biostatistics* 12 (4): 610–23. https://doi.org/10.1093/biostatistics/kxq083.

Tibshirani, Julie, Susan Athey, and Stefan Wager. 2020. *Grf: Generalized Random Forests*. https://CRAN.R-project.org/package=grf.

Twala, BETH, MC Jones, and David J Hand. 2008. “Good Methods for Coping with Missing Data in Decision Trees.” *Pattern Recognition Letters* 29 (7): 950–56.

Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek. 2020. “Methods for Correcting Inference Based on Outcomes Predicted by Machine Learning.” *Proceedings of the National Academy of Sciences* 117 (48): 30266–75. https://doi.org/10.1073/pnas.2001238117.

Watson, Gregory L, Colleen E Reid, Michael Jerrett, and Donatello Telesca. 2020. “Prediction & Model Evaluation for Space-Time Data.” *arXiv Preprint arXiv:2012.13867*.

Wiedinmyer, C, SK Akagi, Robert J Yokelson, LK Emmons, JA Al-Saadi, JJ Orlando, and AJ Soja. 2011. “The Fire Inventory from Ncar (FINN): A High Resolution Global Model to Estimate the Emissions from Open Burning.” *Geoscientific Model Development* 4 (3): 625–41.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The Fair Guiding Principles for Scientific Data Management and stewardship.” *Scientific Data* 3 (March): 160018. https://doi.org/10.1038/sdata.2016.18.

Zandbergen, Paul A, Timothy C Hart, Kathryn E Lenzer, and Michael E Camponovo. 2012. “Error Propagation Models to Examine the Effects of Geocoding Quality on Spatial Analysis of Individual-Level Datasets.” *Spatial and Spatio-Temporal Epidemiology* 3 (1): 69–82.

Zoogman, P, X Liu, RM Suleiman, WF Pennington, DE Flittner, JA Al-Saadi, BB Hilton, et al. 2017. “Tropospheric Emissions: Monitoring of Pollution (TEMPO).” *Journal of Quantitative Spectroscopy and Radiative Transfer* 186: 17–39.