XIS-PM2.5: A daily spatiotemporal machine-learning model for PM2.5 in the contiguous United States

Allan Just[1], Kodi Arfer[1], Johnathan Rush[1], Alexei Lyapustin[2], and Itai Kloog[1]

[1]Icahn School of Medicine at Mount Sinai
[2]NASA Goddard Space Flight Center

December 7, 2022

Abstract

Air-pollution monitoring is sparse across most of the United States, so geostatistical models are important for reconstructing concentrations of fine particulate air pollution (PM2.5) for use in health studies. We present XGBoost-IDW Synthesis (XIS), a daily high-resolution PM2.5 machine-learning model covering the contiguous US from 2003 through 2021. XIS uses aerosol optical depth from satellites and a parsimonious set of additional predictors to make predictions at arbitrary points, capturing near-roadway gradients and allowing the estimation of address-level exposures. We built XIS with a computationally tractable workflow for extensibility to future years, and we used weighted evaluation to fairly assess performance in sparsely monitored regions. Averaging across all years in site-level cross-validation, the weighted mean absolute error of predictions (MAE) was 2.13 μg/m3, a substantial improvement over the mean absolute deviation from the median, which was 4.23 μg/m3. Comparing XIS to a leading product from the US Environmental Protection Agency, the Fused Air Quality Surface Using Downscaling (FAQSD), we obtained a 22% reduction in MAE. We also found a stronger relationship between PM2.5 and social vulnerability with XIS than with the FAQSD. Thus, XIS has potential for reconstructing environmental exposures, and its predictions have applications in environmental justice and human health.

Hosted file

`essoar.10512861.1.docx` available at https://authorea.com/users/539282/articles/611048-xis-pm2-5-a-daily-spatiotemporal-machine-learning-model-for-pm2-5-in-the-contiguous-united-states

XIS-$PM_{2.5}$: A daily spatiotemporal machine-learning model for $PM_{2.5}$ in the contiguous United States

Allan C. Just[1*], Kodi B. Arfer[1], Johnathan Rush[1], Alexei Lyapustin[2], Itai Kloog[1,3]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[3]The Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

Corresponding Author: allan.just@mssm.edu
Address: Allan Just, One Gustave L. Levy Place, Box 1057, New York, NY 10029 USA

Abstract

Air-pollution monitoring is sparse across most of the United States, so geostatistical models are important for reconstructing concentrations of fine particulate air pollution ($PM_{2.5}$) for use in health studies. We present XGBoost-IDW Synthesis (XIS), a daily high-resolution $PM_{2.5}$ machine-learning model covering the contiguous US from 2003 through 2021. XIS uses aerosol optical depth from satellites and a parsimonious set of additional predictors to make predictions at arbitrary points, capturing near-roadway gradients and allowing the estimation of address-level exposures. We built XIS with a computationally tractable workflow for extensibility to future years, and we used weighted evaluation to fairly assess performance in sparsely monitored regions. Averaging across all years in site-level cross-validation, the weighted mean absolute error of predictions (MAE) was 2.13 $\mu g/m^3$, a substantial improvement over the mean absolute deviation from the median, which was 4.23 $\mu g/m^3$. Comparing XIS to a leading product from the US Environmental Protection Agency, the Fused Air Quality Surface Using Downscaling (FAQSD), we obtained a 22% reduction in MAE. We also found a stronger relationship between $PM_{2.5}$ and social vulnerability with XIS than with the FAQSD. Thus, XIS

has potential for reconstructing environmental exposures, and its predictions have applications in environmental justice and human health.

Keywords

- XGBoost
- aerosol optical depth
- absolute error
- air pollution and health
- environmental justice

Synopsis

Improved estimates of air-pollution concentration where people live will improve future analyses on the health impacts of air pollution and exposure disparities across the United States.

Introduction

Particulate-matter air pollution comprises a mixture of solid and liquid particles that are suspended in the air. Concentrations of fine particulate matter ($PM_{2.5}$; having a diameter of less than 2.5 μm) are widely monitored and studied due to associations of short- and long-term exposure to $PM_{2.5}$ with disease[1]. Long-term ambient $PM_{2.5}$ was the leading environmental risk factor and ranked sixth among all modifiable risk factors in the 2019 Global Burden of Disease Study[2], with additional health impacts attributable to short-term $PM_{2.5}$ exposure.

A primary difficulty for such health studies is the geographically sparse monitoring of $PM_{2.5}$, especially over large areas with complex emissions patterns, such as the contiguous United States (CONUS). Regulatory monitoring networks can provide high temporal resolution, with hourly or daily samples[3], but sparse spatial coverage can lead to substantial measurement error on the exposure side of epidemiological analyses[4]. Hence, in place of merely assigning

cases the $PM_{2.5}$ concentration measured at the nearest monitor, researchers increasingly use a variety of methods to model and interpolate $PM_{2.5}$, from chemical-transport models[5] to land-use regression[6] to machine-learning approaches (reviewed in Diao et al)[7]. Even then, however, $PM_{2.5}$ epidemiology has focused on urban areas, where monitoring is most intense[8,9]. This is a critical limitation, since many people across CONUS live in rural or suburban areas, which have a different source profile of air pollution. Especially necessary are evaluation metrics that give rural areas appropriate weight, instead of mostly reflecting performance in intensely monitored areas[10,11].

Machine-learning models for $PM_{2.5}$ typically use predictors of sources (e.g., roadways), topography, and meteorological conditions that relate to $PM_{2.5}$ concentrations. The advent of remotely sensed Earth observations have led many to include satellite aerosol optical depth (AOD) into modeling efforts, including those from our group[12]. Since AOD is a quantitative estimate of the amount of light absorbed or scattered by suspended particles along the vertical atmospheric column, it is a useful proxy for surface $PM_{2.5}$ concentrations[13]. We have developed multiple AOD-based models at various geographical scales (state, region, and country) integrating satellite data with land use regression predictors using linear mixed models to calibrate the satellite-to-surface relationship[14–16].

There is a clear trend for recent $PM_{2.5}$ models to use methods from machine learning, including random forests[14,17,18], gradient boosting[10,19], neural networks[20], and heterogeneous ensembles[21]. While these more flexible models can improve predictive accuracy, we (and others) have shown that without adequate care for the structure of the data, they are prone to overfitting, and data leakage (inadvertent inclusion of test data in model training) can optimistically bias assessments of model performance[10]. Importantly, a flexible model will appear to be much more

accurate when evaluated without consideration of the spatial structure of the underlying phenomenon, and this effect on apparent performance is evidence of overfitting[10]. Another issue with machine learning is the temptation to build huge models that demand extensive computational resources, run slowly, and use hundreds of predictors that would all need to be updated for future years. Such models sound very impressive, but are difficult to scale, update, and reproduce, and are not necessarily more accurate than more conservative models.

We aimed to develop a new computationally efficient national model based on integrating a machine-learning method, namely extreme gradient boosting (XGBoost)[10,22], with inverse-distance weighting (IDW). Predictors included a parsimonious set of satellite, land-use, meteorological, and topographical variables. We call our general modeling framework, which can be used to model not only $PM_{2.5}$ but also other environmental variables such as air temperature, XGBoost-IDW Synthesis (XIS); the adaptation of XIS for $PM_{2.5}$ is named XIS-$PM_{2.5}$. Our model covers CONUS for each day of 2003 through 2021, with planned updates as new data become available. The strengths of XIS include an ungridded core that can make predictions at arbitrary points in the study region; good computational efficiency, with a geospatial data pipeline that dramatically cuts down on time and processing requirements; recency, with pre-planned ability to update the model for new years; a weighted evaluation scheme that fairly considers model performance across the entire region; and interpretability of results. We provide detailed evaluation of predictive performance in site-wise cross-validation, with stratification by year and climate region. We use interpretable machine-learning approaches to better understand the predictor-to-$PM_{2.5}$ relationships learned by our model. We also compare our model with the EPA FAQSD daily tract-level $PM_{2.5}$ predictions, highlighting the value of

point-based predictions. An environmental-justice application shows that XIS, versus EPA FAQSD, produces substantially different estimates of exposure disparities across CONUS.

Method

Study area and time period

We modeled $PM_{2.5}$ for each year from 2003 to 2021. Our study area was CONUS excluding some large water bodies, as defined by the US Census's state-wise cartographic boundary files[23]; it covers 7,798,188 km^2 in all. For some analyses, we show results divided into the 9 standard NOAA climate regions[24,25]. We treat Washington, DC, as part of the Northeast region.

Although XIS represents time as discrete days, in Central Standard Time (UTC−6), it does not discretize space into a grid. Rather, it represents locations as floating-point longitude–latitude pairs.

Data

Particulate Matter

The outcome variable for XIS-$PM_{2.5}$, fine particulate matter mass concentration, is represented as $PM_{2.5}$ measurements recorded in the Environmental Protection Agency's Air Quality System (AQS)[3]. We included monitors using the federal reference method or federal equivalent methods (parameter code 88101) as well as other monitors reporting "acceptable $PM_{2.5}$" (parameter code 88502), including mass concentrations from speciation networks. Overall, 38% of the observations we used came from "acceptable" instruments, and 18% of monitoring sites only became available for this analysis because they include such an instrument. We filtered the outcome as follows:

1. Select observations inside the study area that are based on 24-hour measures.

2. Drop observations with an event type of "Excluded"; use the corresponding "Included" observations instead.
3. Group observations into sites on the basis of longitude and latitude (disregarding AQS site identifiers).
4. Handle each observation of negative $PM_{2.5}$ by either setting it to 0 or discarding it entirely. A negative observation is discarded if it is more than 1 standard deviation (SD) away from the mean of other observations at the same site within the past 3 days and the next 3 days.
5. Reduce the data to at most one observation per site and day. Rank the observations as follows: prefer the designated primary monitor if there is one; prefer parameter code 88101 to 88502; prefer integrated 24-hour measures over 24-hour block averages; prefer lower AQSIDs; and prefer lower parameter occurrence codes.
6. Buddy-check the observations: for each site and day, compute the mean of the observations at all other sites within 30 km, weighted by the inverse square of the distance. If there is at least one such buddy and the interpolation differs from the interpolated-to observation by 20 $\mu g/m^3$ or more, flag the observation for removal. Once all observations have been checked, remove the flagged observations.

Across all years, the result comprised 5,445,462 station-days of observations, 10,655 of which (1 in 511) had the value 0 $\mu g/m^3$.

Predictors

XIS-$PM_{2.5}$ uses the following 22 variables as predictors. Further details of data processing are given below.

- Longitude and latitude

- The integer day of the year
- An IDW feature, which is an interpolation of same-day particulate measurements at sites within 500 km, weighted by the cube of the distance (thus, the IDW exponent is 3)
- Daily AOD at 470 nm, from the MAIAC algorithm for Terra and Aqua (one variable per satellite)[26]
- Daily modeled surface $PM_{2.5}$ concentrations, from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2)[27]
- Monthly vegetation, quantified as the enhanced vegetation index from Aqua[28]
- The daily height of the planetary boundary layer (PBL) from the 5th generation reanalysis of the global climate dataset (ERA5)[29]
- The distance from the nearest fire on the same day, using fire locations from Aqua and Terra[30]
- The distance from the nearest primary road, using the US Census's 2019 national road geodatabase[31]
- Two variables for surface imperviousness (from the National Land Cover Database[32]): one for the imperviousness at a single 30-m grid cell and one for the Gaussian-filtered imperviousness in a 1-km square around the query point
- Population density, from the Gridded Population of the World[33]
- Elevation, from the US Geological Survey's 3D Elevation Program[34]
- Hilliness, or local relative topography, quantified as the multi-scale topographic dissection index computed from elevation[35]

- 6 meteorological variables from the North American Land Data Assimilation System-2 (NLDAS-2): temperature, specific humidity, air pressure, zonal wind speed, meridional wind speed, and precipitation[36]

Additional data processing

The population-density product we used is available in intervals of 5 years, and the imperviousness product is available for irregularly spaced years. For each year of our data, we used the latest update of these products that was not in the future; for example, there is one Gridded Population of the World dataset for 2000 and another for 2005, so we used the 2000 dataset for our 2004 analyses and the 2005 dataset for our 2005 analyses.

ERA5 PBL height from the European Centre for Medium-Range Weather Forecasts (ECMWF) was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store[29]. PBL height, MERRA-2 surface $PM_{2.5}$ concentrations (calculated using the formula from[37]), and NLDAS-2 meteorological variables are available on an hourly basis, so we computed the mean for each day in Central Standard Time.

Elevation data for CONUS were aggregated from 1-arcsecond (~30 m) to 300-m resolution for computational speed. Hilliness was constructed from this aggregated raster as a multi-scale topographic dissection index (using window sizes of 3, 6, 9, 12, and 15 km) following the topoclimatic temperature modeling of Oyler et al.[35].

Where multiple overpasses from the same MODIS instrument occurred over the same 1-km grid cell per day, we selected the non-missing AOD with the lowest associated theoretical AOD uncertainty (based on surface brightness) and restricted to "best quality" or "AOD within +-2km from the coastline" or "Land, research quality" based on the field `AOD_QA`[38].

Models

XIS uses a machine-learning algorithm called extreme gradient boosting (XGBoost)[22]. XGBoost grows a forest of regression trees, fitting each tree to the error of prior trees, and applies several kinds of regularization, allowing it to strike a balance between flexibility, avoidance of overfitting, and computation time. XGBoost also automatically handles missing values: for each split, it chooses a default direction to use when the split variable is missing. Hence, XIS does not need to separately impute missing predictors.

Although the dependent variable for XIS-$PM_{2.5}$ is fine particulate concentration, we don't provide this directly to XGBoost. Instead, we compute the IDW interpolations first, and XGBoost models the observed concentration minus the IDW. For prediction, the IDW is added back. This strategy serves two purposes: it allows XIS to benefit from our prior knowledge that IDW is by far the most important feature, and it produces smoother predictions than raw XGBoost output, which comes in discrete steps because of the use of trees. The IDW is still provided to XGBoost as a feature in case XGBoost can benefit by altering the prediction based on its value. Lastly, although extremely rare, all negative predictions are set to 0, because real particulate concentrations are always nonnegative.

To tune XGBoost's hyperparameters, which control the learning process and model complexity, we first chose 50 hyperparameter vectors at random with a maximin Latin-hypercube algorithm[39] to ensure broad coverage of the hyperparameter space. We conducted cross-validation (see below) on each of two years of our data, 2004 and 2018, with each of these hyperparameter vectors. We used only two years to ensure we had plenty of data to test the models on that we had not already used for tuning hyperparameters. On the basis of computation time and performance in both years and across regions, we chose these hyperparameter values:

`nrounds` 500, `max_depth` 6, `colsample_bytree` 0.5, `eta` 0.073, `gamma` 0.0093, `lambda` 130, `alpha` 0.0012.

Evaluation

Predictive models are often evaluated in terms of squared error, leading to the SD as a metric of variability and root mean square error (RMSE) as a metric of performance. However, $PM_{2.5}$ observations are strongly positively skewed: they are all nonnegative and most values are small, while a minority are very large. For example, among the 350,216 AQS observations from 2019, even after the aforementioned filtering steps, the quartiles were 4, 6.2, 9.2 μg/m^3, and the .95 quantile was 15.6 μg/m^3, but the 50 greatest values ranged from 54 to 113 μg/m^3. Squared error would emphasize performance for this handful of large values, incentivizing models to create sufficiently large predictions to support large values while overestimating the majority of values. Thus, we quantified performance with mean absolute error (MAE), and baseline variability with mean absolute deviation from the median (MAD). In the same way that the SD equals the RMSE that is obtained by predicting every value with the mean, the MAD equals the MAE that is obtained by predicting every value with the median. For XGBoost's objective function, we used log-cosh error, an everywhere twice-differentiable approximation to absolute error, defined by $f(x) = \ln \cosh x$.

We estimated MAE in new data with ten-fold site-wise cross-validation, as follows. In each year, we randomly partitioned all sites with at least one observation into ten equally sized folds. We separately fit XIS on each set of nine folds and made predictions to the held-out fold. During cross-validation, we computed IDW interpolations while holding out sites from both the test fold and the fold of the interpolated-to site.

Were we to take the raw cross-validated MAE as our measure of model performance, a problem would arise due to the spatial distribution of sites. AQS sites are spread unevenly throughout the CONUS, with high concentrations of sites in some places and only a few sites in others. Hence, the unweighted MAE would emphasize performance wherever there happens to be more sites. We weighted observations (in evaluation, but not in training) so as to give equal weight to each unit of spacetime covered by XIS. The method was, for each day and region, to draw a Voronoi diagram[40] for all sites with an observation, and use the areas of the Voronoi tiles, in km^2, as weights for the observations; thus, observations that were relatively isolated were assigned greater weight.

Results

Cross-validation

Table 1 shows the weighted MAD, MAE, and bias (mean signed error) for each year of cross-validated predictions. Each MAE is one or more $\mu g/m^3$ lower than its corresponding MAD, showing that XIS has meaningful predictive ability. Averaging across all years, the mean MAD is 4.23 $\mu g/m^3$ and the mean MAE is 2.13 $\mu g/m^3$. The MAD decreases over the years, and so does the MAE, albeit not as fast, so in later years, accuracy is better but the improvement over baseline is not as impressive. The consistent negative bias is likely due to the skewed distribution of observations, which has a minority of large values. Figure 1 shows the same weighted-MAD and weighted-MAE metrics, but computed separately for each NOAA region. Overall, MAEs are less variable between regions and over time than the MADs are. Averaging across all years, we have the lowest mean MAE in the South region (1.69 $\mu g/m^3$), and the greatest in the Northwest (2.84 $\mu g/m^3$) and West (3.28 $\mu g/m^3$), where there is also the greatest variation in $PM_{2.5}$.

Table 1: Results from each yearly cross-validation, in μg/m^3.

Year	Observations	Sites	MAD	MAE	Bias
2003	181,665	1,309	5.20	2.27	-0.29
2004	197,176	1,262	5.08	2.22	-0.16
2005	215,484	1,326	5.49	2.27	-0.07
2006	217,863	1,291	4.82	2.20	-0.12
2007	248,887	1,282	5.14	2.30	-0.16
2008	256,187	1,283	4.59	2.21	-0.08
2009	274,383	1,309	4.13	2.11	-0.19
2010	288,247	1,298	4.22	2.13	-0.17
2011	285,769	1,255	4.30	2.24	-0.17
2012	295,172	1,245	3.88	2.21	-0.31
2013	302,008	1,248	3.92	2.11	-0.21
2014	308,797	1,271	3.80	2.09	-0.28
2015	317,884	1,288	3.76	2.05	-0.23
2016	321,951	1,261	3.26	1.93	-0.24
2017	332,150	1,274	3.72	2.11	-0.26
2018	342,160	1,283	3.66	1.99	-0.30
2019	350,216	1,280	3.18	1.73	-0.20
2020	353,320	1,262	3.92	2.09	-0.28
2021	356,143	1,274	4.22	2.15	-0.32

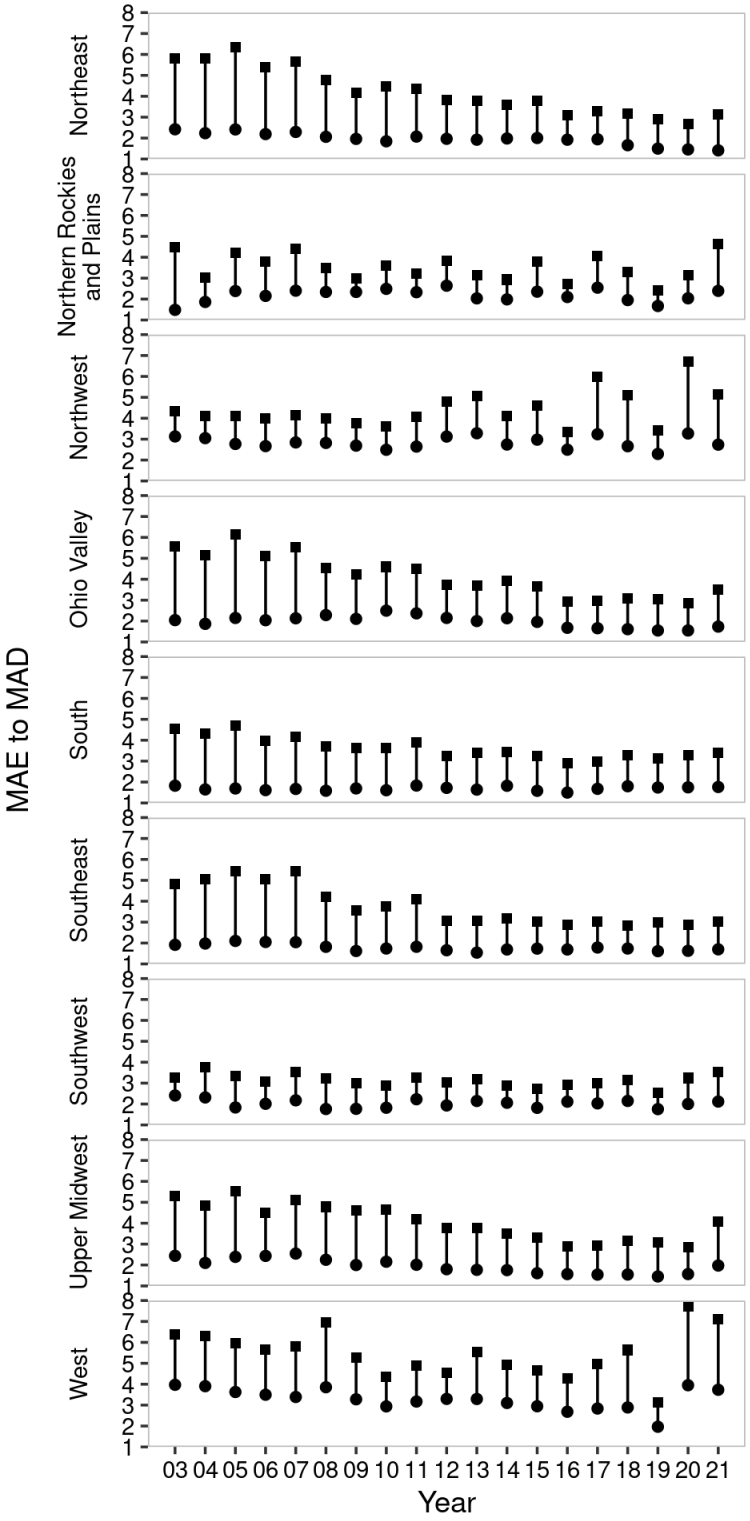

Figure 1: Weighted MAD (squares) and MAE (circles) from cross-validation for each region and year.

Across all years, cross-validation produced 2,385 predictions of 0 μg/m^3 (1 in 2,283 out of all predictions).

See the Supporting Information for additional results including a stratification by season and the cross-validation performance at isolated sites.

Feature contributions

To examine how individual predictors relate to XIS's predictions, we show SHapley Additive exPlanations (SHAPs)[41]. SHAPs can be interpreted analogously to the terms of a linear-regression model: a SHAP of +2.5 for a given predictor and case means that the model attributes a +2.5 increase in its prediction for that case to that predictor. We generated SHAPs for predictions corresponding to each observation while it was held out in cross-validation.

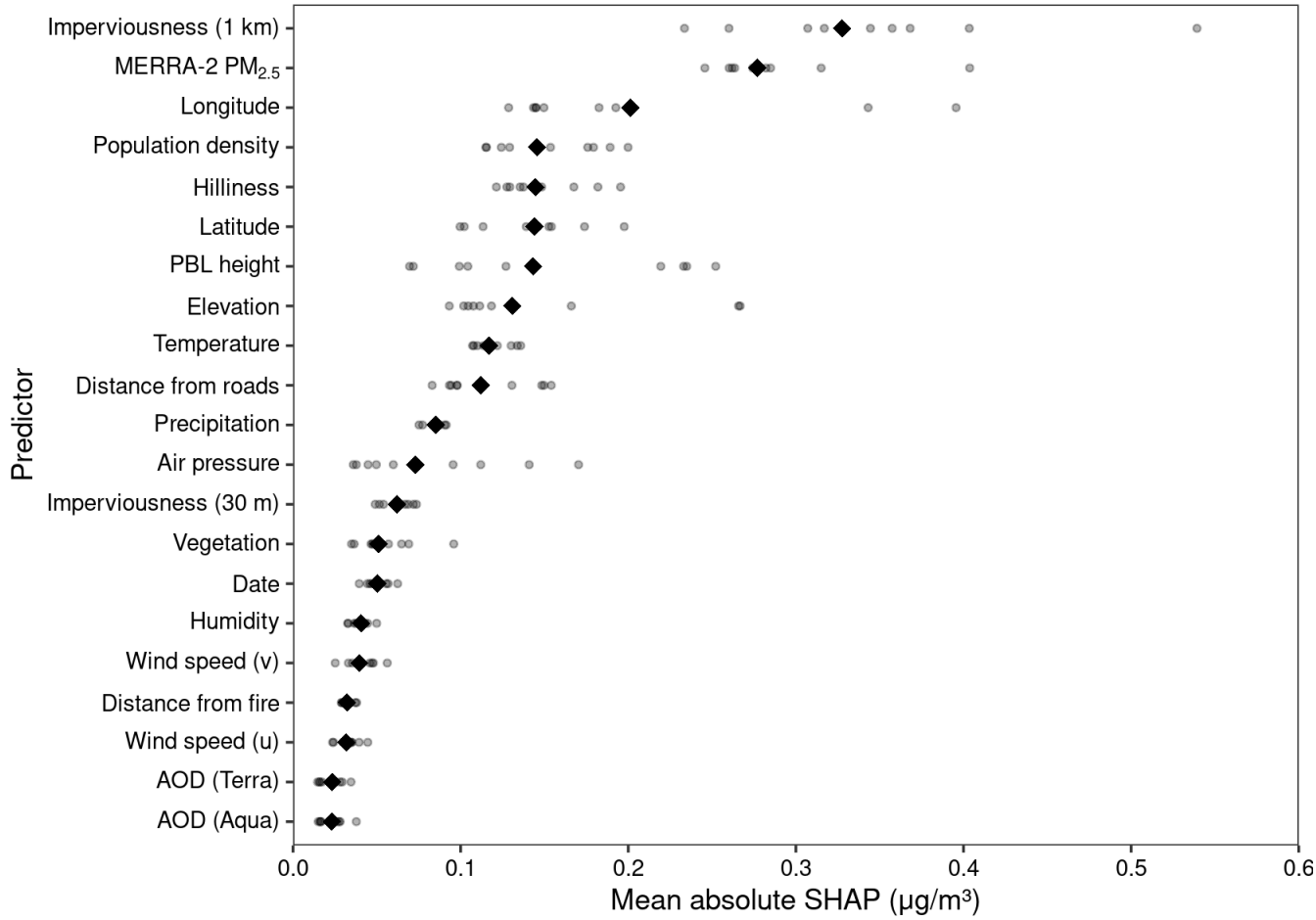

Figure 2: Mean absolute SHAP of each predictor in 2010 (the IDW feature, which has much greater absolute SHAP than everything else, is omitted). Small dots show per-region means. Diamonds show overall means.

Figure 2 shows per-feature mean absolute SHAPs for one year, which can be interpreted as showing the typical impact of the feature on the prediction. Except for the IDW feature, which has a mean absolute SHAP of 9.3 $\mu g/m^3$ (when computed accounting for its initial inclusion outside of XGBoost), each feature has a relatively small contribution, effectively constituting fine adjustments to the IDW interpolation. About half of the features have a mean absolute SHAP greater than 0.1 $\mu g/m^3$.

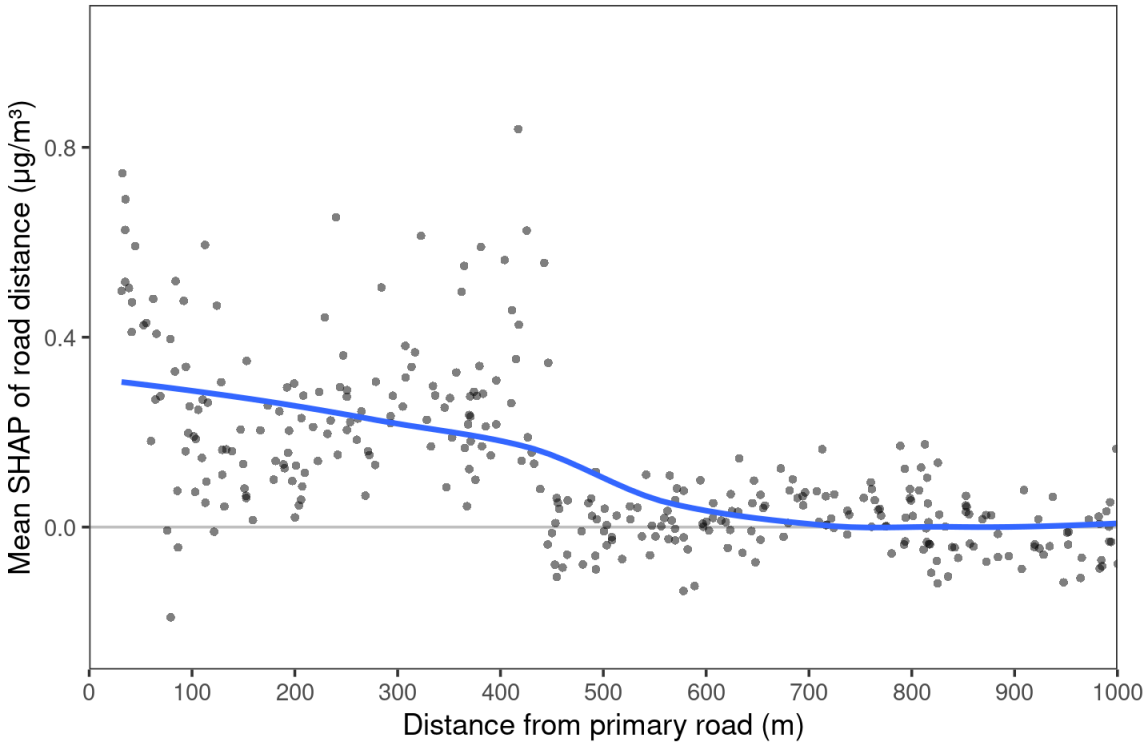

Figure 3: SHAP of road distance as a function of road distance in 2010. A trendline is fit with locally estimated scatterplot smoothing (LOESS).

Figure 3 plots the mean SHAP of the road-distance feature for each AQS site (restricted to those sites within 1 km of a major roadway). As one would expect, low distances are

associated with a higher concentration of $PM_{2.5}$. SHAPs shrink towards 0 as the distance increases, indicating that our model finds the distance from the nearest road less predictively useful as it increases. See the Supporting Information for analyses demonstrating the relation of hilliness with SHAPs and the contribution of MERRA-2 as a function of site isolation.

Comparison with a downscaler

The EPA provides a $PM_{2.5}$ product with one prediction per day and per US Census tract called the Fused Air Quality Surface Using Downscaling (FAQSD), which is currently available through 2018[42,43]. The FAQSD uses a subset of AQS sites for a bias adjustment of surfaces generated from a 12-km resolution Community Multiscale Air Quality (CMAQ) model. We sought a direct comparison of XIS performance to the tract-level output of the FAQSD, taking advantage of the clear delineation of which AQS sites were and were not used by FAQSD[43].

For each year, we denoted all observations from sites that the FAQSD used for training (identified by AQS ID in the FAQSD input files) as the training set, and all remaining observations as the test set. To get an FAQSD prediction for each test observation, we used the nearest FAQSD value (based on tract centroids) on the same day. We retrained XIS yearly using only AQS data from the training set and made two sets of predictions for the test set: one at the same locations as the FAQSD values, and one at the true locations of the test observations. By looking at XIS at both the FAQSD locations (nearest tract centroid) and the true locations of test monitors, we can see how much of the improvement of XIS over FAQSD is due to the improvement in spatial resolution from using point-based predictions.

Table 2 shows for each year the weighted MAD among the test set and the weighted MAEs of the three kinds of predictions. XIS achieved substantially lower MAE than the FAQSD despite how the FAQSD uses the CMAQ and XIS doesn't. There is substantial improvement

when the model is changed from FAQSD to XIS without changing prediction locations, and another, generally smaller improvement when XIS makes predictions for the true locations. Averaging across years, XIS achieves a 16% reduction in MAE compared to FAQSD without changing prediction locations, and a 22% reduction for true locations. Similarly, also averaging across years, XIS achieves a 37% reduction in absolute yearly bias compared to FAQSD without changing prediction locations, and a 57% reduction for true locations.

Table 2: Results of the comparison between XIS and the FAQSD, in $\mu g/m^3$. The observation and site counts include only the test set. We only analyze as far as 2018 because that's the latest year available for the FAQSD.

Year	Observations	Sites	MAD	FAQSD		XIS, tract centroids		XIS, true locations	
				MAE	Bias	MAE	Bias	MAE	Bias
2003	28,342	208	5.09	3.14	0.78	2.53	0.09	2.30	-0.08
2004	38,863	238	4.68	3.14	1.14	2.52	0.26	2.30	0.25
2005	54,222	283	4.70	3.38	1.60	2.44	0.43	2.16	0.30
2006	63,832	315	4.54	3.01	1.18	2.44	0.58	2.16	0.40
2007	76,641	338	4.70	2.97	0.72	2.52	0.30	2.27	0.16
2008	83,365	360	4.21	2.95	0.93	2.39	0.36	2.16	0.19
2009	95,946	405	3.88	3.03	1.07	2.41	0.30	2.25	0.22
2010	107,284	433	3.97	2.77	0.66	2.35	0.19	2.19	0.12
2011	121,353	486	4.16	2.88	0.28	2.44	-0.19	2.24	-0.19
2012	129,135	496	3.79	2.68	0.17	2.40	-0.42	2.21	-0.31
2013	132,387	504	3.83	2.70	0.13	2.36	-0.08	2.16	-0.12
2014	80,717	341	3.72	2.61	0.74	2.30	-0.10	2.12	-0.11
2015	78,731	342	3.66	2.58	0.99	2.11	0.04	2.02	0.05
2016	76,454	319	3.19	2.38	0.65	2.09	-0.07	1.90	-0.06
2017	76,273	316	4.06	2.78	0.17	2.54	-0.36	2.49	-0.10
2018	76,220	310	3.77	2.42	0.33	2.22	-0.46	2.10	-0.30

New predictions

Figure 4 shows the predicted $PM_{2.5}$ concentrations throughout the study region for one year. Figure 5 shows one day in the greater New York City area; here, we plot approximately one prediction per 187 m.

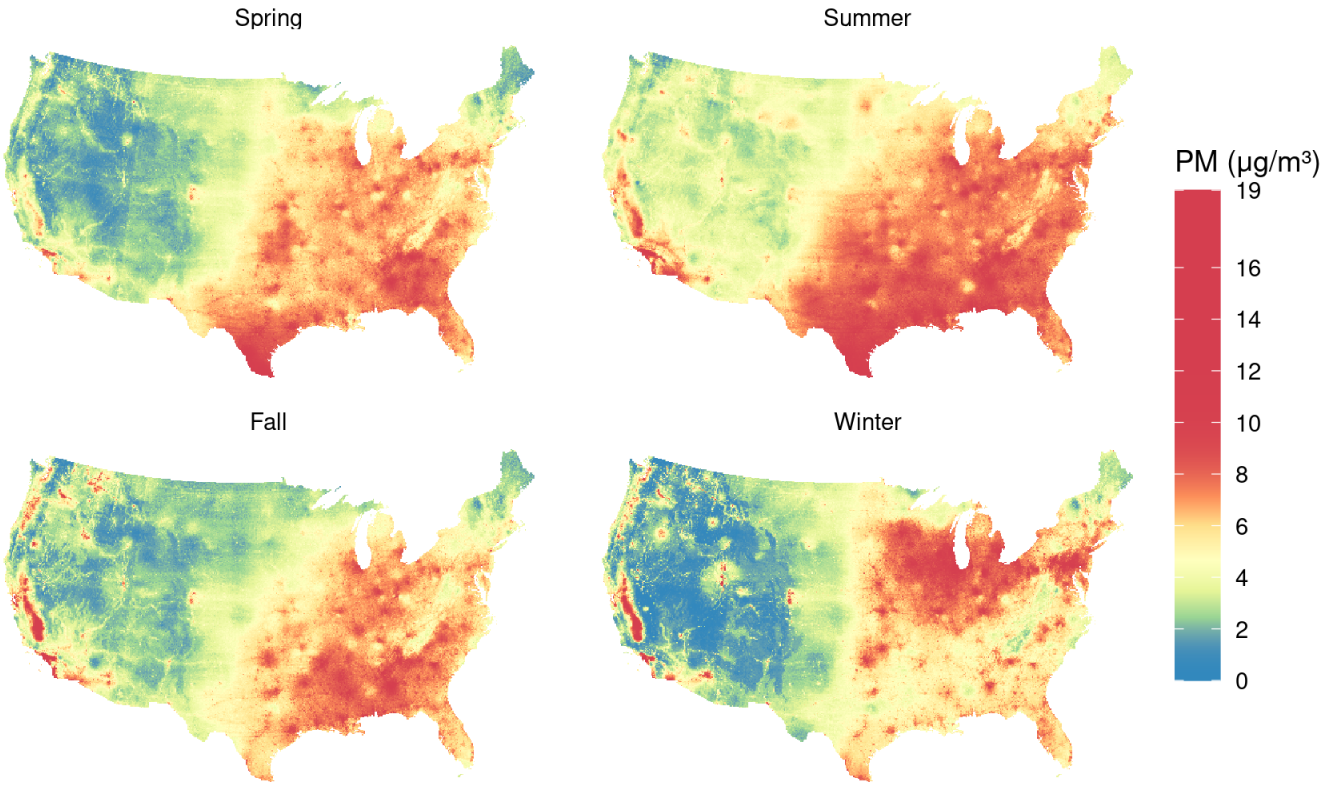

Figure 4: Mean predicted $PM_{2.5}$ across the study area from 1 Dec 2018 through 30 Nov 2019, grouped by season. We use December from 2018 instead of 2019 so as to plot a contiguous winter. Colors are scaled according to quantiles, such that color changes more rapidly around values that are more common in the map.

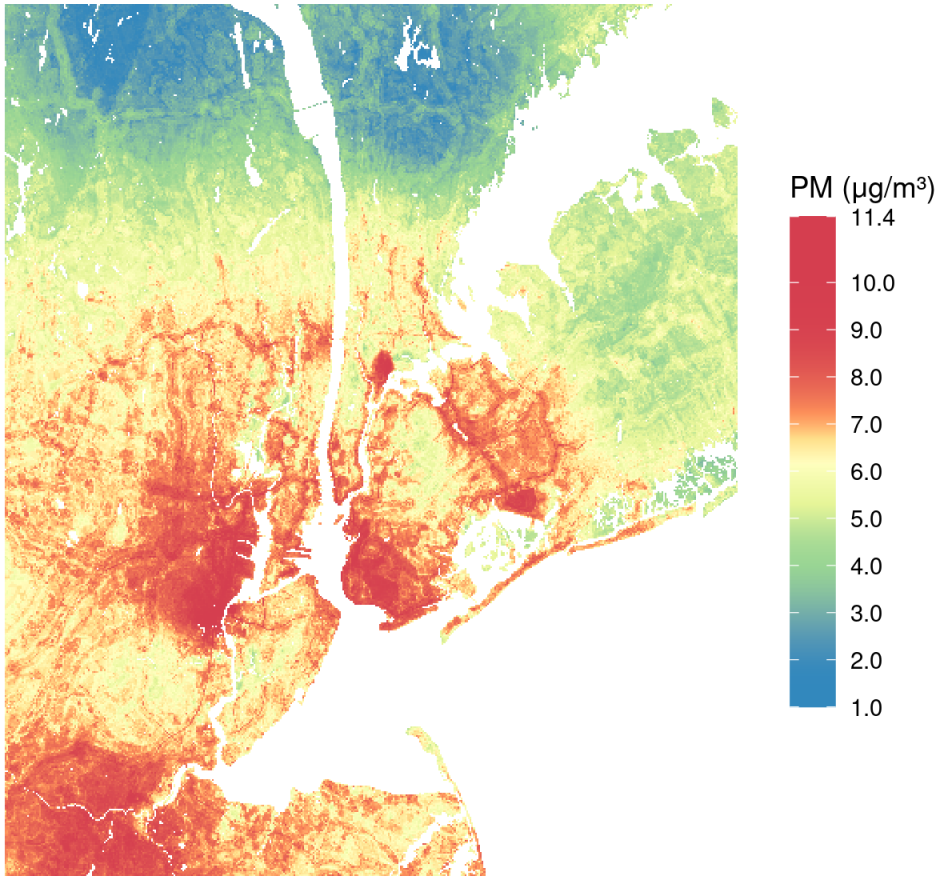

Figure 5: Predicted $PM_{2.5}$ on 10 Jul 2021 in the New York City area. We chose this date by computing the mean $PM_{2.5}$ at all stations in this area for each day in 2021, then taking the median day.

Social vulnerability

We examined how predicted $PM_{2.5}$ at the centroids of 71,619 Census tracts in CONUS, averaged across all days of 2018, related to the CDC's per-tract Social Vulnerability Index (SVI)[44]. 2018 was the latest available year for both FAQSD and SVI. SVI scores range from 0 (least vulnerable) to 1 (most vulnerable). We fit two linear mixed models, one with $PM_{2.5}$ estimates from XIS as the outcome and one with estimates from FAQSD. The models had a fixed effect for vulnerability as well as per-county random intercepts and slopes of vulnerability

(modeled as correlated). The fixed effect of vulnerability was estimated as 0.081 $\mu g/m^3$ (95% CI [0.056, 0.105]) for FAQSD and 0.655 $\mu g/m^3$ (95% CI [0.606, 0.703]) for XIS.

Discussion

We modeled daily $PM_{2.5}$ across CONUS from 2003 through 2021. Our model, XIS, uses a streamlined geospatial processing and machine-learning pipeline to facilitate regular updates with new data. XIS is trained on a broad information base of ground concentrations from the AQS, comprising both federal reference method/federal equivalent method monitors as well as other acceptable $PM_{2.5}$ monitors. While there are many potential uses for this exposure model, particularly in epidemiology and environmental justice, we focus our discussion on metrics of predictive accuracy, comparison with a leading EPA product (the FAQSD), and the interpretation of individual predictors.

We took special care in our evaluation of predictive accuracy. Our site-wise cross-validation scheme, in which we split data spatially before computing the IDW predictor or fitting XGBoost, can estimate accuracy without being optimistically biased by data leakage or overfitting. Thus our performance metrics were evaluated at sites for which the model has seen no prior ground truth, indicating what the performance would be when, for example, estimating exposure at a person's home where there isn't already a monitor. We used MAE as our primary metric, to suitably reflect overall accuracy in the presence of a small number of extreme $PM_{2.5}$ concentrations, and we used a log-cosh objective function to help minimize MAE. Given the goal of covering sparsely monitored suburban and rural regions, we weighted the MAE to reflect the land area covered rather than the number of monitors. The result was overall good accuracy, substantially better than the MAD. Accuracy increased, and the gap between the MAD and MAE shrank, in later years as ambient $PM_{2.5}$ decreased. We also examined unweighted MAE at

isolated monitors to further assess XIS's performance in sparsely monitored regions. Here we again found good accuracy, comparable to that of our overall weighted MAEs.

Our workflow is reproducible and computationally tractable, allowing us to rerun XIS in order to interrogate performance. For example, we trained a version of XIS using only the AQS sites that the FAQSD is trained on, allowing us to fairly and comprehensively compare XIS to the FAQSD. This comparison demonstrated that XIS had a substantially lower MAE in every available year, even when fit without a substantial subset of ground monitors used in our full model. Had XIS required hundreds of predictors and thousands of processor cores, this kind of rerun would have been impractical. Rather, we chose our predictors parsimoniously, and we searched for hyperparameters in a way that was more efficient than exhaustive.

In XIS, as in our prior XGBoost-based models[10,45], we used SHAP to quantify the contribution of individual predictors. SHAPs show that the IDW interpolation of monitor observations is the greatest contributor to predictions at withheld monitors. The IDW interpolation serves as a base prediction (modified by XGBoost) and reuses distance matrices for speed. As an example of XIS's interpretability using SHAP, Figure 3 shows that the positive effect of being near a primary road smoothly decreases with distance, particularly within 500 m. This figure also shows the value of XIS's point-based design, incorporating both raster data and continuous fields, in contrast to our previous 1-km gridded models[10]. Such spatial precision was further justified in the comparison with the FAQSD when we found that using the exact locations of test monitors, compared to tract centroids, increased accuracy and decreased bias (Table 2).

For an environmental-justice application with relevance to human health, we examined the association between social vulnerability and annual $PM_{2.5}$ concentrations at Census tracts throughout CONUS. We see a striking difference between models in estimates of exposure

disparities: XIS's estimate of the relationship between vulnerability and $PM_{2.5}$ is nearly an order of magnitude greater than the FAQSD's. Thus XIS reveals meaningful exposure disparities that would be greatly attenuated by using the FAQSD.

XIS starts in 2003, when MODIS instruments on two satellites (Aqua and Terra) became available, and is updated through 2021. Pairing this long duration (19 years) with such recency is important for health studies in a rapidly changing era. Few other national $PM_{2.5}$ models for use in health studies are regularly updated. The FAQSD is a notable example, although its dependence on the National Emissions Inventory, which is updated every three years, means that the FAQSD is typically several years out of date. We have developed our geospatial processing workflow to ingest new data (including new types of predictors) and make periodic updates so that we can continue to generate timely exposure estimates. Future developments may incorporate near-real time data streams such as EPA AirNOW and the near-real-time MAIAC AOD[46], allowing us to generate predictions up to the past few days and model rapidly evolving air-quality and health conditions. Similarly, we have developed XIS to be adapted for other environmental exposures. Ongoing efforts are generating a suite of complementary exposure estimates (e.g., air temperature and humidity) that are synergistically useful with these $PM_{2.5}$ estimates for health studies.

A limitation of XIS-$PM_{2.5}$ is that it only models mass concentrations. The toxicity of particulates is also related to particulate composition, which XIS does not address. Furthermore, since XIS uses the AQS as ground truth, our ability to update XIS is limited by the AQS release schedule. For example, EPA first released complete 2021 AQS summary files in June 2022. Finally, it remains possible that adding some of the complexity we have avoided, such as hundreds of additional predictors or elaborate ensembles of models, would increase XIS's

accuracy. Diminishing returns, however, means that a great deal of XIS's agility and tractability could be lost for a marginal improvement in accuracy.

In summary, we present a new exposure model, XIS, which generates point-based daily $PM_{2.5}$ exposures across the conterminous United States 2003-2021. This model is intended to generate ambient exposure estimates in cohort and registry-based epidemiological and exposure disparity studies in order to advance the evidence basis of the health impacts of chronic and acute exposure to $PM_{2.5}$.

Acknowledgments

Research reported in this publication was supported by the Environmental influences on Child Health Outcomes (ECHO) program, Office of The Director, National Institutes of Health, under Award Numbers U2C OD023375, UH3 OD023337, and an ECHO Opportunities and Infrastructure Fund award to ACJ, as well as National Institutes of Health grants R01 ES031295, P30 ES023515, and U54 TR003213.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

(1) US Environmental Protection Agency. *Supplement to the 2019 Integrated Science Assessment for Particulate Matter (Final Report, 2022)*. https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=354490 (accessed 2022-06-23).
(2) Health Effects Institute. State of Global Air 2020. HEI Boston 2020.
(3) *EPA AirData*. https://aqs.epa.gov/aqsweb/airdata/download_files.html (accessed 2022-06-23).
(4) Zeger, S. L.; Thomas, D.; Dominici, F.; Samet, J. M.; Schwartz, J.; Dockery, D.; Cohen, A. Exposure Measurement Error in Time-Series Studies of Air Pollution: Concepts and Consequences. *Environ. Health Perspect.* **2000**, *108* (5), 419–426. https://doi.org/10.1289/ehp.00108419.
(5) van Donkelaar, A.; Martin, R. V.; Brauer, M.; Kahn, R.; Levy, R.; Verduzco, C.; Villeneuve, P. J. Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application. *Environ. Health Perspect.* **2010**, *118* (6), 847–855. https://doi.org/10.1289/ehp.0901623.

(6) Hoek, G.; Beelen, R.; De Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A Review of Land-Use Regression Models to Assess Spatial Variation of Outdoor Air Pollution. *Atmos. Environ.* **2008**, *42* (33), 7561–7578.
(7) Diao, M.; Holloway, T.; Choi, S.; O'Neill, S. M.; Al-Hamdan, M. Z.; Van Donkelaar, A.; Martin, R. V.; Jin, X.; Fiore, A. M.; Henze, D. K.; Lacey, F.; Kinney, P. L.; Freedman, F.; Larkin, N. K.; Zou, Y.; Kelly, J. T.; Vaidyanathan, A. Methods, Availability, and Applications of PM $_{2.5}$ Exposure Estimates Derived from Ground Measurements, Satellite, and Atmospheric Models. *J. Air Waste Manag. Assoc.* **2019**, *69* (12), 1391–1414. https://doi.org/10.1080/10962247.2019.1668498.
(8) Hoek, G.; Krishnan, R. M.; Beelen, R.; Peters, A.; Ostro, B.; Brunekreef, B.; Kaufman, J. D. Long-Term Air Pollution Exposure and Cardio- Respiratory Mortality: A Review. *Environ. Health* **2013**, *12* (1), 43. https://doi.org/10.1186/1476-069X-12-43.
(9) Pope, C. A.; Dockery, D. W. Health Effects of Fine Particulate Air Pollution: Lines That Connect. *J. Air Waste Manag. Assoc.* **2006**, *56* (6), 709–742. https://doi.org/10.1080/10473289.2006.10464485.
(10) Just, A. C.; Arfer, K. B.; Rush, J.; Dorman, M.; Shtein, A.; Lyapustin, A.; Kloog, I. Advancing Methodologies for Applying Machine Learning and Evaluating Spatiotemporal Models of Fine Particulate Matter (Pm2.5) Using Satellite Data over Large Regions. *Atmos. Environ.* **2020**, *239*, 117649. https://doi.org/10.1016/j.atmosenv.2020.117649.
(11) Carrión, D.; Arfer, K. B.; Rush, J.; Dorman, M.; Rowland, S. T.; Kioumourtzoglou, M.-A.; Kloog, I.; Just, A. C. A 1-Km Hourly Air-Temperature Model for 13 Northeastern U.S. States Using Remotely Sensed and Ground-Based Measurements. *Environ. Res.* **2021**, *200*, 111477. https://doi.org/10.1016/j.envres.2021.111477.
(12) Kloog, I.; Koutrakis, P.; Coull, B. A.; Lee, H. J.; Schwartz, J. Assessing Temporally and Spatially Resolved Pm2.5 Exposures for Epidemiological Studies Using Satellite Aerosol Optical Depth Measurements. *Atmos. Environ.* **2011**, *45* (35), 6267–6275. https://doi.org/10.1016/j.atmosenv.2011.08.066.
(13) Just, A. C.; De Carli, M. M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I. Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA. *Remote Sens (Basel)* **2018**, *10* (5), 803. https://doi.org/10.3390/rs10050803.
(14) Stafoggia, M.; Schwartz, J.; Badaloni, C.; Bellander, T.; Alessandrini, E.; Cattani, G.; de'Donato, F.; Gaeta, A.; Leone, G.; Lyapustin, A.; Sorek-Hamer, M.; de Hoogh, K.; Di, Q.; Forastiere, F.; Kloog, I. Estimation of Daily Pm10 Concentrations in Italy (2006–2012) Using Finely Resolved Satellite Data, Land Use Variables and Meteorology. *Environ. Int.* **2017**, *99*, 234–244. https://doi.org/10.1016/j.envint.2016.11.024.
(15) Just, A. C.; Wright, R. O.; Schwartz, J.; Coull, B. A.; Baccarelli, A. A.; Tellez-Rojo, M. M.; Moody, E.; Wang, Y.; Lyapustin, A.; Kloog, I. Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily PM $_{2.5}$ Geographical Distribution in Mexico City. *Environ. Sci. Technol.* **2015**, *49* (14), 8576–8584. https://doi.org/10.1021/acs.est.5b00859.
(16) Hough, I.; Sarafian, R.; Shtein, A.; Zhou, B.; Lepeule, J.; Kloog, I. Gaussian Markov Random Fields Improve Ensemble Predictions of Daily 1 Km Pm2. 5 and Pm10 Across France. *Atmos. Environ.* **2021**, *264*, 118693.
(17) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y. Estimating PM $_{2.5}$ Concentrations in the Conterminous United States Using the Random

Forest Approach. *Environ. Sci. Technol.* **2017**, *51* (12), 6936–6944. https://doi.org/10.1021/acs.est.7b01210.
(18) Shtein, A.; Kloog, I.; Schwartz, J.; Silibello, C.; Michelozzi, P.; Gariazzo, C.; Viegi, G.; Forastiere, F.; Karnieli, A.; Just, A. C. Estimating Daily Pm2. 5 and Pm10 over Italy Using an Ensemble Model. *Environ. Sci. Technol.* **2019**, *54* (1), 120–128.
(19) Reid, C. E.; Jerrett, M.; Petersen, M. L.; Pfister, G. G.; Morefield, P. E.; Tager, I. B.; Raffuse, S. M.; Balmes, J. R. Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. *Environ. Sci. Technol.* **2015**, *49* (6), 3887–3896. https://doi.org/10.1021/es505846r.
(20) Di, Q.; Koutrakis, P.; Schwartz, J. A Hybrid Prediction Model for Pm2.5 Mass and Components Using a Chemical Transport Model and Land Use Regression. *Atmos. Environ.* **2016**, *131*, 390–399. https://doi.org/10.1016/j.atmosenv.2016.02.002.
(21) Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Mickley, L. J.; Schwartz, J. An Ensemble-Based Model of Pm2.5 Concentration Across the Contiguous United States with High Spatiotemporal Resolution. *Environ. Int.* **2019**, *130*, 104909. https://doi.org/10.1016/j.envint.2019.104909.
(22) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System, 2016. https://doi.org/10.1145/2939672.2939785.
(23) US Census. *Cartographic Boundary Files*. https://www2.census.gov/geo/tiger/GENZ2019/description.pdf (accessed 2020-10-07).
(24) NOAA. *U.S. Climate Regions | Monitoring References | National Centers for Environmental Information (NCEI)*. https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php (accessed 2020-10-07).
(25) Karl, T. R.; Koscielny, A. J. Drought in the United States: 1895–1981. *Journal of Climatology* **1982**, *2* (4), 313–329. https://doi.org/10.1002/joc.3370020402.
(26) Lyapustin, Alexei; Wang, Yujie. Mcd19a2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2g Global 1km SIN Grid V006, 2018. https://doi.org/10.5067/MODIS/MCD19A2.006.
(27) Global Modeling And Assimilation Office. MERRA-2 Tavg1_2d_aer_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Aerosol Diagnostics V5.12.4, 2015. https://doi.org/10.5067/KLICLTZ8EM9D.
(28) Didan, Kamel. MODIS/Aqua Vegetation Indices Monthly L3 Global 1km SIN Grid V061, 2021. https://doi.org/10.5067/MODIS/MYD13A3.061.
(29) Hersbach, H.; Bell, B.; Berrisford, P.; Biavati, G.; Horányi, A.; Muñoz Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Rozum, I. Era5 Hourly Data on Single Levels from 1979 to Present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* **2018**, *10*. https://doi.org/10.24381/cds.adbb2d47.
(30) Giglio, L.; Schroeder, W.; Justice, C. O. The Collection 6 MODIS Active Fire Detection Algorithm and Fire Products. *Remote Sens. Environ.* **2016**, *178*, 31–41.
(31) US Census Bureau, Geography Division. *2019 TIGER/Line Shapefiles*. https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2019&layergroup=Roads (accessed 2022-06-17).
(32) Dewitz, J. National Land Cover Database (NLCD) 2019 Products, 2021. https://doi.org/10.5066/P9KZCM54.

(33) Center For International Earth Science Information Network-CIESIN-Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11, 2018. https://doi.org/10.7927/H4JW8BX5.
(34) US Geological Survey. *1 Arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection*. https://www.sciencebase.gov/catalog/item/4f70aa71e4b058caae3f8de1 (accessed 2022-06-15).
(35) Oyler, J. W.; Ballantyne, A.; Jencso, K.; Sweet, M.; Running, S. W. Creating a Topoclimatic Daily Air Temperature Dataset for the Conterminous United States Using Homogenized Station Data and Remotely Sensed Land Skin Temperature. *Int. J. Climatol* **2015**, *35* (9), 2258–2279. https://doi.org/10.1002/joc.4127.
(36) NASA. *NLDAS-2 Forcing Dataset Information*. https://ldas.gsfc.nasa.gov/nldas/v2/forcing (accessed 2022-06-23).
(37) Global Modeling and Assimilation Office. *MERRA-2 FAQ*. https://web.archive.org/web/2021/https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/FAQ (accessed 2022-06-09).
(38) Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC Algorithm. *Atmos. Meas. Tech.* **2018**, *11* (10), 5741–5765. https://doi.org/10.5194/amt-11-5741-2018.
(39) Carnell, R. *maximinLHS (lhs: Latin Hypercube Samples)*. https://CRAN.R-project.org/package=lhs (accessed 2020-10-07).
(40) Turner, R. *deldir: Delaunay triangulation and Dirichlet (Voronoi) tessellation*. https://CRAN.R-project.org/package=deldir (accessed 2021-01-04).
(41) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles, 2019. http://arxiv.org/abs/1802.03888.
(42) Adam Reff. Bayesian Space-Time Downscaling Fusion Model (Downscaler) - Derived Estimates of Air Quality for 2018. *EPA-454/R-21-003* **2021**, 135.
(43) US EPA, O. *RSIG-Related Downloadable Data Files*. https://www.epa.gov/hesc/rsig-related-downloadable-data-files (accessed 2022-06-07).
(44) Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. *CDC/ATSDR Social Vulnerability Index (SVI) 2018 Database US*. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html (accessed 2022-06-17).
(45) Just, A. C.; Liu, Y.; Sorek-Hamer, M.; Rush, J.; Dorman, M.; Chatfield, R.; Wang, Y.; Lyapustin, A.; Kloog, I. Gradient Boosting Machine Learning to Improve Satellite-Derived Column Water Vapor Measurement Error. *Atmos. Meas. Tech.* **2020**, *13* (9), 4669–4681.
(46) MODIS Land Science Team. MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2g Global 1km SIN Grid, 2019. https://doi.org/10.5067/MODIS/MCD19A2N.NRT.006.

Supporting Information

XIS-$PM_{2.5}$: A daily spatiotemporal machine-learning model for $PM_{2.5}$ in the contiguous United States

Allan C. Just[1*], Kodi B. Arfer[1], Johnathan Rush[1], Alexei Lyapustin[2], Itai Kloog[1,3]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[3]The Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

Corresponding Author: allan.just@mssm.edu
Address: Allan Just, One Gustave L. Levy Place, Box 1057, New York, NY 10029 USA

Table 3: 2019 cross-validation results (μg/m^3) broken down by meteorological season. Results for December are taken from the 2018 model so that a contiguous winter is analyzed.

Season	Observations	Sites	MAD	MAE	Bias
all	349,993	1,284	3.15	1.72	-0.20
Spring	88,936	1,256	2.99	1.55	-0.20
Summer	88,942	1,253	2.81	1.47	-0.13
Fall	87,001	1,245	3.14	1.70	-0.21
Winter	85,114	1,284	3.63	2.16	-0.27

Table 3 shows cross-validation results by season for one year. Compared to the whole one-year period, MAD and MAE are lower in spring and summer and higher in winter.

Table 4: Results from each yearly cross-validation among isolated sites, in μg/m^3.

Year	Observations	Sites	MAD	MAE	Bias
2003	25,913	236	4.67	1.93	-0.03
2004	29,202	231	4.43	2.00	0.07
2005	30,520	225	5.02	2.09	-0.01
2006	34,457	233	4.38	2.08	0.01
2007	37,396	225	4.78	2.23	-0.06
2008	38,129	226	4.24	2.05	-0.20
2009	40,172	223	3.83	1.94	-0.18
2010	44,377	230	3.98	1.98	-0.06
2011	44,935	223	4.18	2.21	-0.25
2012	46,399	222	3.83	2.15	-0.27
2013	49,163	224	3.66	2.10	-0.11
2014	54,287	238	3.59	2.04	-0.26
2015	55,638	242	3.59	2.04	-0.19
2016	57,603	241	3.04	1.86	-0.17
2017	60,450	239	3.49	2.02	-0.13
2018	62,863	239	3.58	1.95	-0.28
2019	63,538	242	3.08	1.73	-0.20
2020	65,667	238	3.49	2.01	-0.22
2021	64,125	242	3.97	2.06	-0.25

In addition to the weighted analyses using all stations, we wished to evaluate performance where ground networks were especially sparse. Thus, Table 4 shows unweighted MAE from cross-validation among the sites that were particularly isolated, defined as being more than 50 km from all other sites available in the same year.

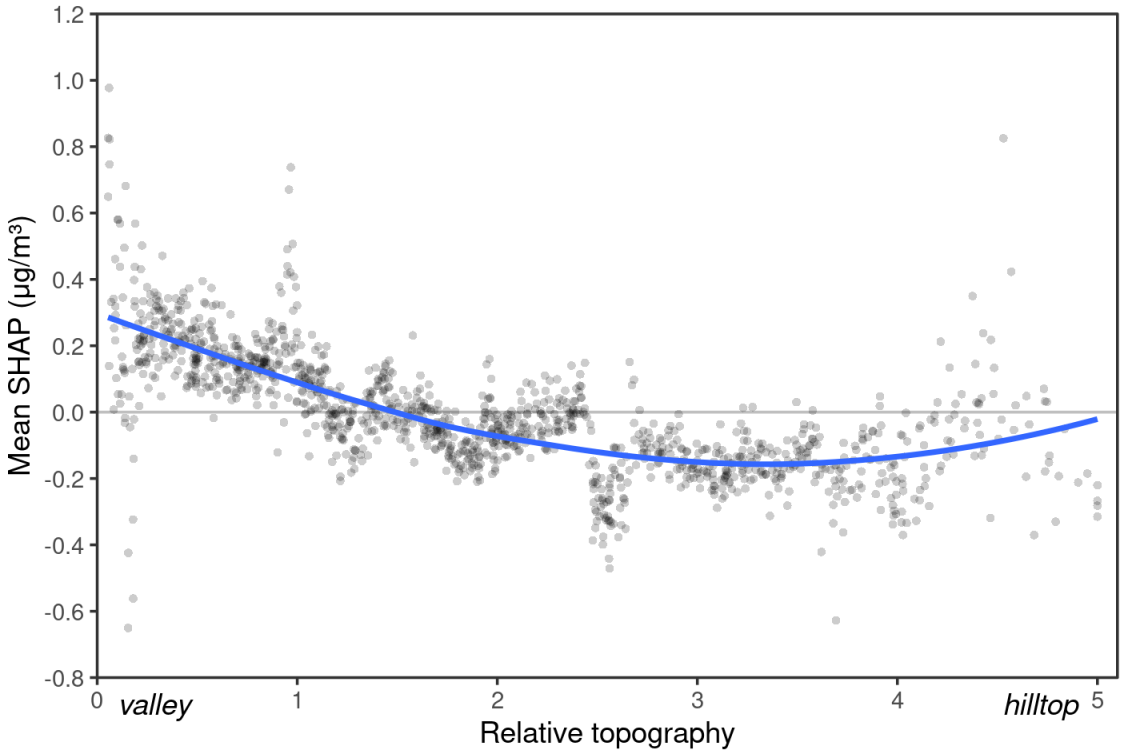

Figure 6: SHAP of hilliness as a function of hilliness.

Figure 6, similar to Figure 3, plots the mean SHAP of the hilliness feature for each site. We see higher average predicted $PM_{2.5}$ at sites in a valley versus on a hill.

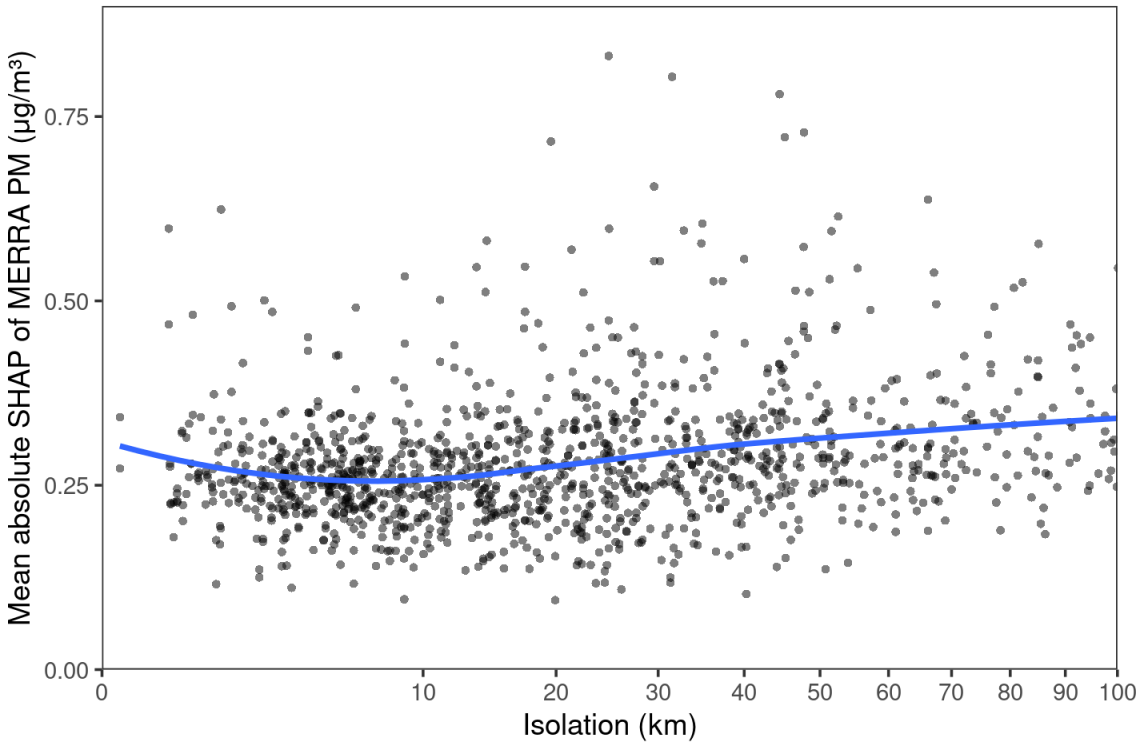

Figure 7: SHAP of modeled surface $PM_{2.5}$ concentrations from MERRA-2 as a function of site isolation. The *x*-axis is on a square-root scale.

Figure 7 is an example of the relationship between the SHAP of one variable and a different quantity. The *y*-axis shows the per-site mean absolute SHAP of MERRA-2 modeled $PM_{2.5}$, but the *x*-axis shows the site's distance from its nearest neighbor; that is, its degree of isolation. Similarly, we examined how the SHAP of the IDW feature varied according to isolation. The per-site mean absolute SHAP for IDW in 2010 was Kendall-correlated -0.17 with the distance to the nearest other site, meaning that the IDW is less influential on predictions for more isolated sites.