HHS Public Access

Author manuscript

J Expo Sci Environ Epidemiol. Author manuscript; available in PMC 2023 March 10.

Published in final edited form as:

J Expo Sci Environ Epidemiol. 2022 November ; 32(6): 917–925. doi:10.1038/s41370-022-00471-4.

Prediction of daily mean and one-hour maximum $PM_{2.5}$ concentrations and applications in Central Mexico using satellite-based machine-learning models

Iván Gutiérrez-Avila[*,1], **Kodi B. Arfer**[1], **Daniel Carrión**[1,2,3], **Johnathan Rush**[1], **Itai Kloog**[1,4], **Aaron R. Naeger**[5], **Michel Grutter**[6], **Víctor Hugo Páramo-Figueroa**[7], **Horacio Riojas-Rodríguez**[8], **Allan C. Just**[1,9]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]Department of Environmental Health Sciences, Yale University School of Public Health, New Haven, CT, USA

[3]Center on Climate Change and Health, Yale University School of Public Health, New Haven, CT, USA

[4]Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

[5]Earth System Science Center, University of Alabama in Huntsville, Huntsville, AL, USA

[6]Instituto de Ciencias de la Atmósfera y Cambio Climático, Universidad Nacional Autónoma de México, Ciudad de México, México

[7]Comisión Ambiental de la Megalópolis, Ciudad de México, México

[8]Dirección de Salud Ambiental, Instituto Nacional de Salud Pública, Cuernavaca Morelos, México

[9]Institute for Exposomic Research, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

[*]**Correspondence**: Iván Gutiérrez-Avila, Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1057, New York, NY 10029, USA. ivan_2c@hotmail.com.

Author contribution Statement

Iván Gutiérrez-Avila: Conceptualization, Methodology, Formal Analysis, Writing–Original Draft, Data Curation. **Kodi B. Arfer:** Conceptualization, Methodology, Formal Analysis, Writing–Original Draft, Data Curation, Software. **Daniel Carrión:** Conceptualization, Writing–Review & Editing, **Johnathan Rush**: Conceptualization, Formal Analysis, Writing–Review & Editing, Data Curation, Software. **Itai Kloog:** Conceptualization, Methodology, Writing–Review & Editing. **Aaron R. Naeger:** Conceptualization, Writing–Review & Editing, **Michel Grutter:** Conceptualization, Writing–Review & Editing, **Víctor Hugo Páramo-Figueroa:** Conceptualization, Writing–Review & Editing, **Horacio Riojas-Rodríguez:** Conceptualization, Writing–Review & Editing, **Allan C. Just**: Conceptualization, Methodology, Supervision, Writing–Review & Editing, Project Administration, Resources.

Ethical Approval

Not applicable

Competing interests

The authors declare that they have no conflicts of interest regarding this study.

Background.—Machine-learning algorithms are becoming popular techniques to predict ambient air $PM_{2.5}$ concentrations at high spatial resolutions (1×1 km) using satellite-based aerosol optical depth (AOD). Most machine-learning models have aimed to predict 24h-averaged $PM_{2.5}$ concentrations (mean $PM_{2.5}$) in high-income regions. Over Mexico, none have been developed to predict subdaily peak levels, such as the maximum daily one-hour concentration (max $PM_{2.5}$).

Objective.—Our goal was to develop a machine-learning model to predict mean $PM_{2.5}$ and max $PM_{2.5}$ concentrations in the Mexico City Metropolitan Area from 2004 through 2019.

Methods.—We present a new modeling approach based on extreme gradient boosting (XGBoost) and inverse-distance weighting that uses AOD, meteorology, and land-use variables. We also investigated applications of our mean $PM_{2.5}$ predictions that can aid local authorities in air-quality management and public-health surveillance, such as the co-occurrence of high $PM_{2.5}$ and heat, compliance with local air-quality standards, and the relationship of $PM_{2.5}$ exposure with social marginalization.

Results.—Our models for mean and max $PM_{2.5}$ exhibited good performance, with overall cross-validated mean absolute errors (MAE) of 3.68 and 9.20 $\mu g/m^3$, respectively, compared to mean absolute deviations from the median (MAD) of 8.55 and 15.64 $\mu g/m^3$. In 2010, everybody in the study region was exposed to unhealthy levels of $PM_{2.5}$. Hotter days had greater $PM_{2.5}$ concentrations. Finally, we found similar exposure to $PM_{2.5}$ across levels of social marginalization.

Keywords

machine-learning model; environmental modeling; particulate matter; remote sensing; air quality management; air pollution

1. Introduction

Fine particulate matter with aerodynamic diameter ≤ 2.5 microns ($PM_{2.5}$) affects more people than any other pollutant, and has been consistently associated with mortality and morbidity from cardiovascular and respiratory causes (1). Over the last decade, epidemiological evidence has related $PM_{2.5}$ to many other health outcomes, such as cardio-metabolic diseases (including diabetes, hypertension, metabolic syndrome), neurological disorders (stroke, dementia, Alzheimer's disease, autism, Parkinson's disease), and perinatal outcomes (premature birth and low birth weight) (2). At the same time, exposure scientists have developed new modeling approaches for air-pollution epidemiology, moving away from the use of data from ground monitors alone. Interest has grown in models using remote-sensing products, particularly aerosol optical depth (AOD) for the prediction of ground level $PM_{2.5}$ concentrations at high spatial resolutions, such as 1 x 1 km. AOD is a measure of the amount of light absorbed and scattered throughout the atmospheric vertical column by the collection of suspended particles (e.g., urban haze, smoke, desert dust, sea salt) in the atmosphere. AOD is related to $PM_{2.5}$ concentrations as measured by ground monitors, but the relationship is complex and depends on a number of other factors (3). Popular approaches to predicting ground-level $PM_{2.5}$ concentration using AOD include chemical-transport models, mixed-effect models, geographically weighted regression, and land-use regression, which use additional $PM_{2.5}$ predictors and modifiers of the $PM_{2.5}$–AOD

relation such as weather and land use (3, 4). Among the most comprehensive efforts to reconstruct ground concentrations of $PM_{2.5}$ is NASA's Global Modeling Initiative (GMI) chemistry transport model integrated with Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2 GMI), which estimates the global distribution of $PM_{2.5}$ mass concentrations with a spatial resolution of 0.5° × 0.625°, and temporal resolution as fine as 1 hour (5, 6).

Predicting ground-based $PM_{2.5}$ from satellite AOD retrievals is difficult. AOD is strongly influenced by particles above the surface layer, which have different characteristics from ground-level particles. Also, AOD retrieval algorithms assume consistent particle size distributions within large regions, such as Mexico and Central America (7). Furthermore, AOD often has gaps in spatial coverage due to clouds, snow, or ice. Thus, researchers must often impute missing AOD (8), and the complex relationship between AOD and $PM_{2.5}$, along with the use of additional $PM_{2.5}$ predictors, has motivated machine-learning approaches such as neural networks, random forests, and gradient boosting (4, 9–12). Given the challenges to develop a single model that fits large heterogeneous regions (e.g. national models), ensemble models combining the outputs from different machine learning algorithms have been used in recent studies (9).

AOD-based $PM_{2.5}$ (AOD-$PM_{2.5}$) models and predictions have allowed epidemiologists to move away from exposure-assessment methods that rely on proximity to sparse ground monitors. With sufficient spatiotemporal resolution, AOD-$PM_{2.5}$ models may further improve exposure assessment in epidemiologic research by picking up the effects of microenvironments. Few AOD-$PM_{2.5}$ models exist for middle-income countries. Our group developed one of the first AOD-$PM_{2.5}$ models using daily Multi-Angle Implementation of Atmospheric Correction (MAIAC) spectral AOD derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on NASA's Aqua satellite at a 1x1 km spatial resolution, along with data from ground monitors, land use, and meteorological features (7). Our previous model for the Mexico City region provides daily $PM_{2.5}$ predictions from 2004–2014, and those predictions have been used in several epidemiologic studies in this region (13). However, model improvements are needed to better characterize the spatiotemporal distribution of $PM_{2.5}$, particularly since the Mexico City Metropolitan Area has undergone considerable urban sprawl. $PM_{2.5}$ in large metropolitan areas affects not only people in the city center but also people in its suburban and rural outskirts (14). People in the outskirts, where air-quality information is scarce, may face disproportionate health risks due to lower socioeconomic status and less access to healthcare. This environmental injustice can be even more pronounced in low- and middle-income regions (15).

AOD-$PM_{2.5}$ models covering large urban areas have great value for epidemiology, but also for public-health surveillance (e.g. quantifying mortality and morbidity attributable to $PM_{2.5}$) (16), environmental regulation (e.g. assessment of compliance with air quality standards) (17), and risk communication (e.g. designing air-quality indices) (18). Furthermore, AOD-$PM_{2.5}$ models can help air-quality administrators to see trends in the spatiotemporal distribution of $PM_{2.5}$, map hotspots in regions with few monitors, identify emissions sources to consider for abatement actions, as well as forecast and surveillance of

air pollution contingencies and wildfires (19). Overall, AOD-$PM_{2.5}$ models can be powerful aids for decision-making.

Most of the satellite-based $PM_{2.5}$ models yield predictions of 24-hour mean concentrations, perhaps driven by traditional approaches in epidemiology that have focused on this exposure metric, which in turn support standards for daily $PM_{2.5}$ levels. There is growing interest in identification of specific sub-daily $PM_{2.5}$ exposures (e.g., peak concentrations) that may trigger the onset of adverse health outcomes and harm vulnerable people. To our knowledge, this is the first model reconstructing sub-daily $PM_{2.5}$ concentrations in Mexico.

In this study, we present a new model based on extreme gradient boosting (XGBoost) and inverse-distance weighting (IDW) that uses satellite and land-use variables to predict daily mean and max $PM_{2.5}$ concentrations in Central Mexico. We use predictions from our models for novel and policy-relevant analyses of the determinants and distribution of population exposures.

2. Methods

We constructed and evaluated two models: one for daily mean $PM_{2.5}$, spanning 2004 through 2019, and one for daily max $PM_{2.5}$ (i.e. the greatest hourly concentration of $PM_{2.5}$ observed each day), spanning 2011 through 2019. We restricted our max $PM_{2.5}$ predictions to 2011 onwards because of greater coverage of ground monitoring stations. Days were defined according to UTC–6, which coincides with the local time of the study region (Mexico's Zona Centro) when daylight-saving time is not in effect (namely, before the first Sunday of April and after the last Sunday of October).

2.1. Study region

We modeled an irregularly shaped area of 6,650 km^2, 127 km in diameter, around Mexico City. The model used a grid of 7,745 square cells, 927 m on a side, in a global sinusoidal projection (the same one used for NASA's MODIS products). This study area and its grid was a subset of that considered in our ambient temperature model for Central Mexico (20). We built the subset by finding the largest connected set of cells in the Valley of Mexico with all cells ≤ 3 km above sea level (Figure 1). The Valley of Mexico is a plateau with a mean elevation of 2,250 m above sea level, and is surrounded on three sides by mountain ranges, preventing the dispersion of air pollutants (21).

2.2. Data

We used $PM_{2.5}$ data from ground monitoring stations organized by the Instituto Nacional de Ecología y Cambio Climático de México (INECC) including records from the Automated Atmospheric Monitoring Network (RAMA) from the Mexico City's Atmospheric Monitoring System (SIMAT, website http://www.aire.cdmx.gob.mx/). We downloaded observations from INECC's website (http://scica.inecc.gob.mx). For each station in the study area and day of $PM_{2.5}$ observations, we computed the mean and max $PM_{2.5}$ among the hourly observations, so long as there were at least 18 hours of observations in the day. Other station-days were discarded. The result was a total of 60,365 station-days from 25 stations for mean $PM_{2.5}$ and 40,819 station-days from the same 25 stations for max

$PM_{2.5}$. The number of days of observations contributed per station ranged from 266 to 5,198 (median 2,030) for mean $PM_{2.5}$, and from 50 to 2,901 (median 1,753) for max $PM_{2.5}$.

Our models used the following 14 predictors:

* Longitude and latitude in degrees

* The date, as an integer count of days

* The IDW mean (exponent 2) of all observations of the same dependent variable (i.e., mean $PM_{2.5}$ or max $PM_{2.5}$) on the given day

* MAIAC AOD from NASA's Terra and Aqua satellites (22), with 1 km spatial resolution, whose local overpass times range from 10:40 to 15:15 and 13:10 to 15:05, respectively. We used the primary MCD19A2 product of AOD at 470 nm.

* $PM_{2.5}$ ($\mu g/m^3$) as predicted by MERRA-2 GMI at the surface level, with ~50 km spatial resolution (6), either the mean of the day's 24 hourly values (for modeling mean $PM_{2.5}$) or the value at 10:00 UTC–6 (for max $PM_{2.5}$)

* Temperature (K), precipitation (mm), and vapor pressure (Pa) from Daymet (23) with 1 km spatial resolution, and the temperature being computed as the mean of Daymet's maximum and minimum temperature

* The height of the planetary boundary layer (m) and meridional and zonal wind speeds (m/s) from the 5^{th} generation reanalysis of the global climate dataset (ERA5) of the European Centre for Medium-Range Weather Forecasts (ECMWF), was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (24), using the mean of the day's 24 hourly values (for mean $PM_{2.5}$) or the value at 10:00 UTC–6 (for max $PM_{2.5}$), with ~30 km spatial resolution

* The density of roads (m/km^2) from OpenStreetMap (25), considering only primary, secondary, residential, and tertiary roads

We selected the midmorning time of day 10:00 UTC–6 in constructing some of the predictors for the max $PM_{2.5}$ model because it was the most frequent hour of greatest daily per-station $PM_{2.5}$ concentration in our sample.

2.3. Model evaluation

We evaluated models with leave-one-station-out cross-validation (CV). There are 25 stations, so for each dependent variable, we fit the model 25 times, leaving out one station from training and then testing the model's predictions on the left-out station. We evaluated models with absolute loss rather than squared loss so as not to overweight the importance of a minority of very high observed concentrations of $PM_{2.5}$. Absolute loss leads to mean absolute error (MAE) as a natural measure of predictive accuracy (in place of root mean square error, RMSE, for squared loss), and mean absolute deviation from the median (MAD) as a measure of baseline variation in place of the standard deviation (SD) for squared loss. Note that R^2, which is often used for model assessment, is defined as a squared-loss metric.

For our study, we compute R^2 as 1 minus the MSE divided by the variance, and we show R^2, RMSE, and SD in tables for completeness, although the models are more properly judged in terms of absolute loss.

When computing the IDW predictor during CV, we excluded the held-out station to avoid data leakage.

2.4. Models

We predicted $PM_{2.5}$ with XGBoost (26), a scheme for fast boosted decision trees. We used a log-cosh objective function to approximate absolute loss. Instead of providing $PM_{2.5}$ as the dependent variable to XGBoost directly, we provided $PM_{2.5}$ minus the IDW interpolation and added the IDW back to XGBoost's predictions. This method partly smooths out the otherwise discrete predictions produced by decision trees. We tuned XGBoost with twofold station-wise CV; during the larger CV discussed above, this CV was nested within each fold. Tuning adjusted four hyperparameters:

* The number of trees, which could be 10, 25, 50, or 100

* The maximum tree depth, which could be 3, 6, or 9

* The learning rate η, which could range from 0.01 to 0.5

* A ridge penalty λ, which could range from 2^{-10} to 2^{10}

We preselected a set of 25 random vectors from this space with a maximin Latin-hypercube sample using the function 'maximinLHS' from the R library 'lhs', version 1.1.3 (27).

Once the outer CV was done, to make new predictions, we trained the two models (one for mean $PM_{2.5}$ and one for max $PM_{2.5}$) on all the data, with one more tuning CV apiece. These final models had the following hyperparameters, obtained from the aforementioned tuning procedure: for mean $PM_{2.5}$, 10 trees, max depth 3, $\eta = 0.047$, $\lambda = 10$; for max $PM_{2.5}$, 25 trees, max depth 9, $\eta = 0.073$, $\lambda = 260$.

2.5. Applications

We present three applications of our $PM_{2.5}$ predictions for the Mexico City Metropolitan Area. We examined co-occurring exposures to high $PM_{2.5}$ concentrations and high temperatures from our published spatiotemporal model (20). Person-time estimates of exposure relied on population density estimates for 2010. We estimated the population density within each of our grid cells using the R package exactextractr (28) to calculate the area-weighted mean of the population density in the intersecting Gridded Population of the World (GPWv4) ~1-km raster cells (29). The GPWv4 used data from the 2010 census in Mexico at the level of Área Geoestadística Básica (AGEBs; the Mexican equivalent of US Census tracts). When comparing exposures to permissible annual limits, we computed "yearly" means as the means of four 3-month means, per the Mexican standard (30). Finally, we examined how AGEB-level $PM_{2.5}$ exposure varied with social marginalization within the study region (31). Every AGEB was assigned the mean $PM_{2.5}$ prediction of all 1x1 km grid cells whose centroids fell within the AGEB polygon.

We used R 4.2.0 (32) with package xgboost 1.4.1.1 (33) for analysis.

3. Results

Overall, the observed $PM_{2.5}$ that we trained and tested on had a median of 23 μg/m^3 (MAD = 8.55, IQR = 14.08) for mean $PM_{2.5}$, and a median of 44 μg/m^3 (MAD = 15.64, IQR = 25.00) for max $PM_{2.5}$.

The model for mean $PM_{2.5}$ achieved a MAE of 3.68 μg/m^3 (compared to a MAD of 8.55 μg/m^3), and the model for max $PM_{2.5}$ achieved a MAE of 9.20 μg/m^3 (compared to a MAD of 15.64 μg/m^3). These differences indicate a substantial improvement in accuracy compared to assigning the median exposure to all places and times throughout the study domain. The much greater MAE for max $PM_{2.5}$ than mean $PM_{2.5}$ is to be expected, because maxima are inherently more difficult to predict than means. Tables 1 and 2 show the performance of these models stratified by year.

We also compared model performance by season: cold dry (spanning November through February), warm dry (March to May), and rainy (June to October) (34). Supplementary Table 1 shows that the largest improvement in prediction accuracy (MAD minus MAE) was observed during the cold dry season for both mean and max $PM_{2.5}$ models, although this season still had the largest MAE.

Supplementary Table 2 shows the Pearson correlations among observed and predicted $PM_{2.5}$ for both models. As would be expected, all four variables are positively related. Predictions are more associated with the kind of observation they are meant to predict than the other kind, but there are also strong correlations between mean and max $PM_{2.5}$.

After making predictions for every grid cell and day with both models, we mapped the per-cell mean $PM_{2.5}$ and max $PM_{2.5}$ averaged over 2019 (Figure 2). Discontinuities in the prediction surfaces evident in our maps are the result of model-based splits selected in the longitude and latitude predictors. Although we also include an IDW interpolation that adds some smoothness, XGBoost selects for the most predictively accurate model. Smoothing our predictions more aggressively could make for more intuitive maps, but would not necessarily improve predictive accuracy. As expected, the highest concentrations (shown in dark purple) are in the center-north and center-east subregions of the Mexico City Metropolitan Area (north and east of Mexico City, respectively), with the highest population density and industrial land use. This pattern is also visible in the max $PM_{2.5}$ map, but is most pronounced in the center-north. The lowest $PM_{2.5}$ concentrations (shown in light purple and yellow) are in the southwest, corresponding to the least populated and the most vegetated subregion.

We examined the per-day ratio (collapsing across all cells) of mean and max $PM_{2.5}$. Supplementary Figure 1 shows this ratio for each day in 2019. Generally, the max is about twice the mean, but the ratio decreases in the first half of the year and increases in the second. During the rainy season (June to October), we examined how the ratio differed between days with and without a mean per-cell precipitation of at least 1 mm, and found little difference: the mean ratio was 2.03 on dry days and 2.13 on rainy days.

With our temperature model (20), we examined the relationship between mean daily $PM_{2.5}$ and mean daily temperature. The Kendall correlation between the two over the whole study period was 0.05, indicating a very weak positive relationship overall. Figure 3 breaks this relationship down by season. It can be observed that the $PM_{2.5}$ concentrations are more stable and remain high during the cold dry season, which has been related to the stable atmospheric conditions and frequent thermal inversions in the study region. For the warm dry season and rainy season, there is a clearer tendency for higher $PM_{2.5}$ concentrations on hotter days.

Considering the 88,399 cell-days in which mean $PM_{2.5}$ exceeded Mexico's permissible daily limit of 41 $\mu g/m^3$ (30), the median temperature was 19.2 °C, somewhat warmer than the median in all other cell-days, 15.9 °C. Considering the 173,170 cell-days with a mean temperature of at least 20 °C, we found substantially higher median $PM_{2.5}$, 30.2 $\mu g/m^3$, than in all other cell-days, 19.7 $\mu g/m^3$.

We used population density from GPWv4 in every prediction cell of the study area to estimate person-days of $PM_{2.5}$ exposure in 2010, referring to Mexico's standards for annual and daily ambient concentrations of $PM_{2.5}$ (30). We compared the exposure estimated by our XGBoost-with-IDW model to that estimated by IDW alone, a $PM_{2.5}$ interpolation technique that has been used for a health-impact assessment in this region (35). The study area contained 20,279,491 people in 2010. According to both our model and the IDW-only model, every single person in the Mexico City Metropolitan Area experienced a yearly mean $PM_{2.5}$ worse than the permissible limit of 10 $\mu g/m^3$. The large majority of people (97%, or more than 99% according to IDW) experienced a yearly mean more than twice the limit. Similarly, all people experienced at least one day with a mean $PM_{2.5}$ worse than the daily permissible limit of 41 $\mu g/m^3$. People experienced a mean of 21.6 (23.7 according to IDW) days exceeding the limit. The total number of exceeded person-days was 439 million (481 million according to IDW). Overall, we find widespread exposure to worse-than-permissible air pollution, although our full model suggests slightly less exposure than an IDW-only model. To show population exposure distributions over time, we also calculated the annual average concentration for each populated grid cell for each year, using more than 45 million model predictions. Figure 4 shows the empirical cumulative distribution functions for these annual concentrations calculated with 2010 census population densities. As observed in Figure 4, there has been an overall reduction in the annual exposure to $PM_{2.5}$ since the earliest years (2004 and 2005); however, there is considerable variability in the estimated annual exposures, with less clear recent trends.

We used an index of social marginalization developed by the Consejo Nacional de Población (CONAPO), which considers access to education and health, housing characteristics, and possession of goods (31), to compare urban marginalization in 2010 to mean $PM_{2.5}$. There were 2,065 AGEBs with available marginalization scores (AGEBs' median area was 0.46 km^2, range 0.014 to 7.4 km^2), with one score per AGEB and year, so we summarized mean $PM_{2.5}$ in 2010 by AGEB. Overall, marginalization and $PM_{2.5}$ were Kendall-correlated 0.024, which is a relationship in the expected direction (i.e., AGEBs with more marginalized populations being exposed to more air pollution), but very weak. Breaking the AGEBs into 0.5-unit groups of marginalization (with one group for marginalization −2 to −1.5, one for

−1.5 to −1, etc.), we find a small range of mean per-group $PM_{2.5}$, from 21.78 to 22.56 $\mu g/m^3$.

4. Discussion

We constructed and validated models to predict mean and max $PM_{2.5}$ in the Mexico City Metropolitan Area, and examined potential applications in air-pollution epidemiology and air-quality management. Our machine-learning-based model is the first of its kind in Mexico, although previously, our team used mixed-effects models with AOD to predict mean $PM_{2.5}$ in this region (13). Also new is our consideration of max $PM_{2.5}$, an exposure metric that is becoming relevant to address subdaily health effects from peak exposures to $PM_{2.5}$ (36). Overall, our models exhibited good performance, with prediction errors that decreased over time, as the number of ground monitoring stations increased. Our per-year R^2 for mean $PM_{2.5}$ ranged from 0.64 to 0.86, similar to the R^2 values for our team's XGBoost model in the Northeastern US, which ranged from 0.64 to 0.80 (11). Our new modeling approach could be extended to other regions with low or intermediate density of ground monitoring stations.

Recently the ensemble model framework has become a popular approach to combine $PM_{2.5}$ estimates from different machine-learning models, mostly in data-rich regions where ensemble models have utilized tens to over 100 predictors (9). The implementation of ensemble models in sparsely monitored regions like the Mexico City Metropolitan Area would be challenging because it typically requires withholding more data in order to construct model weights. Despite their potential benefits, the incremental performance from ensemble models compared to single machine-learning algorithms has been reported as small, especially when the base learners perform well (e.g. $R^2 > 0.7$), and the same predictors are involved (9). Overall, the performance of our XGBoost model to predict mean $PM_{2.5}$ was good, and similar to the performance of other tree-based models using a single learner (10), or ensembles using XGBoost (37) or not (9) as one of their learners. Boosted trees (fitting trees sequentially to the residual error of the prior ensemble) typically outperform the independent trees in random forests. XGBoost's multiple forms of regularization help to avoid overfitting and achieve high accuracy and it is often a best-in-class predictive algorithm with smaller datasets (26, 38).

$PM_{2.5}$ predictions from AOD-$PM_{2.5}$ models have been used in epidemiology to reduce exposure measurement error, but may also be useful for applications such as air-quality management, particularly in sparsely monitored regions (19). Figure 2 shows wide variation in both $PM_{2.5}$ metrics across the Mexico City Metropolitan Area. More $PM_{2.5}$ has historically been observed in the center-north and center-east (in the densely populated limits between Mexico City and the State of Mexico), where there are substantial emissions from industry and traffic (39). Our $PM_{2.5}$ predictions allowed us to assess exposure to $PM_{2.5}$ in the entire Mexico City Metropolitan Area, unlike previous studies that could only partly cover this region with data from ground monitoring stations alone (35). The estimated annual mean concentrations from our model exceeded the current annual $PM_{2.5}$ Mexican permissible limits across the entire study region, supporting previous results pointing out that despite significant improvements in the air quality of Mexico City for PM_{10} and ozone

since the 1990s, there remain substantial obstacles to reducing emissions of $PM_{2.5}$ and its precursors (40). The use of our spatiotemporally resolved $PM_{2.5}$ predictions should improve future health impact assessments and support targeted exposure reduction strategies in this region (41).

Seasonally, there is a well-defined pattern of higher $PM_{2.5}$ concentrations during the two dry seasons (Nov-May), due to frequent thermal inversions and stable atmospheric conditions, which favors the accumulation of $PM_{2.5}$. The lowest $PM_{2.5}$ concentrations occur during the rainy season (June-Oct), due to wet deposition (42). We hypothesized that the observed pattern in the daily ratios of mean and max $PM_{2.5}$ (Supplementary Figure 1) reflects the influence of seasonal meteorological conditions. We checked whether higher ratios observed during the rainy season could be explained by precipitation, since late-afternoon showers can reduce $PM_{2.5}$ (42). However, we found that days with at least 1 mm of daily precipitation had only a 5% greater ratio than other days. Evidence from cities at high elevations (>2000 m above sea level) has shown that relative humidity interacts with precipitation and $PM_{2.5}$ emission sources to increase or decrease $PM_{2.5}$ concentrations (43). Increasing relative humidity can raise $PM_{2.5}$ concentration depending on the $PM_{2.5}$ composition and hygroscopic growth ability, especially in traffic-heavy residential areas where only strong rain events (e.g. precipitation >9 mm) are effective in removing $PM_{2.5}$ from the atmosphere. In industrial areas, high relative humidity conditions are more important to decrease $PM_{2.5}$ concentrations, regardless of rain events. Weak rain episodes (e.g. precipitation <1 mm), can also increase $PM_{2.5}$ concentrations by worsening traffic in rush hours and reducing combustion efficiency (43). It is possible that the ratios of mean and max $PM_{2.5}$ observed in Supplementary Figure 1 are produced by the interaction of precipitation, humidity, and $PM_{2.5}$ sources in the study region.

In the context of climate change, it is important to characterize the increasingly common joint occurrence of extreme air pollution and extreme temperatures (44). We found that while $PM_{2.5}$ and temperature are only weakly related overall, higher $PM_{2.5}$ concentrations tended to occur on warmer days, particularly in the rainy season (Figure 3), and conversely, days with mean temperatures of at least 20° C had a substantially worse median $PM_{2.5}$ concentration than cooler days. It has been reported that co-occurring extreme $PM_{2.5}$ and extreme temperatures may increase the acute risk of illness (45), and that the influence of $PM_{2.5}$ on mortality rates may be stronger in warmer cities (46). Previous studies in the Mexico City Metropolitan Area have suggested stronger associations with mortality on days with high $PM_{2.5}$ and extreme temperatures (47), but they may have estimated effects imprecisely, given their citywide approach for estimating exposure. Our $PM_{2.5}$ predictions can improve exposure assessment and air-pollution epidemiology, including studies addressing the interactive effects of $PM_{2.5}$ with temperature.

To put into perspective the human cost of $PM_{2.5}$ exposure, we found that in 2010, every person in the study region was exposed to unhealthy air quality according to the Mexican standards for annual (10 $\mu g/m^3$) and daily (41 $\mu g/m^3$) concentrations, which are several times the recently enacted World Health Organization Guidelines of 5 and 15 $\mu g/m^3$, respectively (48). Overall, in 2010 the population of the study region experienced a mean of nearly three weeks of $PM_{2.5}$ above the current daily Mexican permissible limit.

For epidemiologic research, the distribution of continuous exposures is more relevant for health studies than the dichotomous assessment or duration of compliance with a particular standard. The annual empirical cumulative distributions for all inhabited areas in the study region in Figure 4 are a summary of the population distribution of our exposure estimates that is suitable for assessment of long-term ambient $PM_{2.5}$ exposures and related chronic health effects.

Concentrations of $PM_{2.5}$ measured in a single monitoring station are used to represent the pollution conditions over large spatial domains (up to tens of kilometers) for a specific amount of time, such as one day or one year. However, $PM_{2.5}$ levels can be rapidly influenced by local sources, increasing not only concentrations between monitoring sites, but also the risks of acute health effects. A distinguishing feature of our model is that we also generated a sub-daily metric of $PM_{2.5}$ concentrations, namely, max $PM_{2.5}$ at a 1-km resolution. There are not yet any air-quality standards for sub-daily $PM_{2.5}$ concentrations, but new research into the health impacts from such exposures could eventually support new standards (49, 50). The US Environmental Protection Agency states that “Because a focus on annual average and 24-hour average $PM_{2.5}$ concentrations could mask sub-daily patterns, and because some health studies examine PM exposure durations shorter than 24-hours, it is useful to understand the broader distribution of sub-daily $PM_{2.5}$ concentrations” (36). Because it’s more difficult to reconstruct extrema (e.g., max $PM_{2.5}$) than measures of central tendency (e.g., mean $PM_{2.5}$), future work on estimating health impacts from max $PM_{2.5}$ could particularly benefit from estimating and propagating prediction uncertainty into downstream analyses (51).

Our comparison of $PM_{2.5}$ exposure across levels of social marginalization did not suggest meaningful differences between groups. However, the 2010 Mexican index of social marginalization was only available for urban AGEBs: those with a total population of more than 2,500. Without data for rural AGEBs or irregular settlements, it is naturally more difficult to assess the influence of socioeconomic status. Since the methods employed in the construction of the Mexican index of social marginalization have changed over time, it would be difficult to analyze multiple years and make sense of the differences between them. One study found that in Mexico City in 2015, per-AGEB deprivation was positively associated with PM_{10}, but negatively associated with ozone (14). In this region, PM_{10} concentrations are highly influenced by local emissions from point and area sources (mainly unpaved roads), which may explain why PM_{10} was associated with deprivation. However, $PM_{2.5}$ is strongly influenced by mobile sources, and most of the $PM_{2.5}$ concentrations are secondary aerosols that can travel far from their emission sources, leading to homogeneous $PM_{2.5}$ concentrations (39). AGEBs are the smallest geographic units with information on marginalization scores, and homogeneous socioeconomic characteristics are expected within AGEBs. Nonetheless, it is also possible that socioeconomic variation exists within AGEBs given the large variability in their size, which might affect correct classification of unequally exposed groups.

Despite the good performance of our models throughout the study period, we observed seasonal differences in their performance, which have also been reported in other studies (9, 10, 12). This suggests that seasonal differences are less a property of our model than a

property of the data. The implications of these seasonal differences on the accuracy of $PM_{2.5}$ predictions for exposure assessment in epidemiologic research should be addressed in future studies. Also, as in any other $PM_{2.5}$ prediction strategy, our models depend on the location of ground monitors, which may be not representative of the entire study area; therefore, error in $PM_{2.5}$ prediction can arise especially in remote locations.

A particular limitation of our max $PM_{2.5}$ model arises from the limited temporal resolution in the AOD data. Each satellite passes over the Central Mexico region only once during each period of daylight, possibly missing sudden episodes of intense $PM_{2.5}$. However, the overpass time of the Terra satellite is similar to the daily peak of $PM_{2.5}$ according to ground monitoring stations, so in general, Terra AOD should be representative of max $PM_{2.5}$. Future work will utilize AOD data from the Advanced Baseline Imager (ABI) aboard NOAA's Geostationary Operational Environmental Satellite - R Series (GOES-16 and GOES-17) with temporal resolution as high as 5 minutes over Mexico City. Synergistic AOD products developed from the ABI and upcoming NASA geostationary Tropospheric Emissions: Monitoring of Pollution (TEMPO) mission, planned for launch in 2023, will further enhance capabilities to predict and monitor $PM_{2.5}$ concentrations in the region. TEMPO will advance exposure science in North America, particularly by providing hourly observations of aerosols and gaseous pollutants for supporting air-pollution models (52, 53).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

Data used in the preparation of this article were obtained from the Sistema Nacional de Información de Calidad del Aire (SINAICA) from the Instituto Nacional de Ecología y Cambio Climático (INECC). We also obtained ground monitoring stations' geographic location from INECC's Coordinatión General de Contaminatión y Salud Ambiental and advice from Maria Guadalupe Tzintzun Cervantes, Sub-Directorate of Air Quality. The results contain modified Copernicus Climate Change Service information 2022. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

Funding information

This work was supported by grants from the National Institutes of Health (NIH): R01 ES013744, R01 ES014930, R01 ES021357, R01 ES031295, R01 ES032242, R24 ES028522, P30 ES023515. D.C. was supported by T32 HD049311.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are thoroughly described in the article and links are provided in the reference section. Derived data supporting the findings of this study are available from the corresponding author on request.

References

1. World Health Organization. Ambient (outdoor) air pollution [Internet]. 2021 [cited 2021 Dec 28]. Available from: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

2. Gu J, Shi Y, Zhu Y, Chen N, Wang H, Zhang Z, et al. Ambient air pollution and cause-specific risk of hospital admission in China: A nationwide time-series study. PLoS Med. 2020 Aug;17(8):e1003188. [PubMed: 32760064]
3. Chu Y, Liu Y, Li X, Liu Z, Lu H, Lu Y, et al. A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. Atmosphere. 2016 Oct 14;7(10):129.
4. Sorek-Hamer M, Chatfield R, Liu Y. Review: Strategies for using satellite-based products in modeling PM2.5 and short-term pollution episodes. Environment International. 2020;144:106057. [PubMed: 32889481]
5. Strode SA, Ziemke JR, Oman LD, Lamsal LN, Olsen MA, Liu J. Global changes in the diurnal cycle of surface ozone. Atmos Environ. 2019 Feb 15;199:323–33.
6. MERRA-2 GMI [Internet]. [cited 2022 Feb 15]. Available from: https://acd-ext.gsfc.nasa.gov/Projects/GEOSCCM/MERRA2GMI/
7. Lyapustin A, Wang Y, Korkin S, Huang D. MODIS Collection 6 MAIAC algorithm. Atmos Meas Tech. 2018 Oct 18;11(10):5741–65.
8. Li L, Franklin M, Girguis M, Lurmann F, Wu J, Pavlovic N, et al. Spatiotemporal Imputation of MAIAC AOD Using Deep Learning with Downscaling. Remote Sens Environ. 2020 Feb;237:111584. [PubMed: 32158056]
9. Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution [Internet]. Vol. 130, Environment International. 2019. p. 104909. Available from: 10.1016/j.envint.2019.104909
10. Schneider R, Vicedo-Cabrera AM, Sera F, Masselot P, Stafoggia M, de Hoogh K, et al. A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain. Remote Sens (Basel). 2020 Nov;12(22):3803. [PubMed: 33408882]
11. Just AC, Arfer KB, Rush J, Dorman M, Shtein A, Lyapustin A, et al. 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 Oct 15;239:117649.
12. Wei J, Huang W, Li Z, Xue W, Peng Y, Sun L, et al. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach [Internet] Vol. 231, Remote Sensing of Environment. 2019. p. 111221. Available from: 10.1016/j.rse.2019.111221
13. Just AC, Wright RO, Schwartz J, Coull BA, Baccarelli AA, Tellez-Rojo MM, et al. Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily PM2.5 Geographical Distribution in Mexico City. Environ Sci Technol. 2015 Jul 21;49(14):8576–84. [PubMed: 26061488]
14. Lome-Hurtado A, Touza-Montero J, White PCL. Environmental Injustice in Mexico City: A Spatial Quantile Approach. Exposure and Health. 2020 Jun 1;12(2):265–79.
15. Bravo MA, Ebisu K, Dominici F, Wang Y, Peng RD, Bell ML. Airborne fine particles and risk of hospital admissions for understudied populations: Effects by urbanicity and short-term cumulative exposures in 708 U.s. counties. Environ Health Perspect. 2016 Sep 20;125(4):594–601. [PubMed: 27649448]
16. Southerland VA, Brauer M, Mohegh A, Hammer MS, van Donkelaar A, Martin RV, et al. Global urban temporal trends in fine particulate matter (PM) and attributable health burdens: estimates from global datasets. Lancet Planet Health [Internet]. 2022 Jan 5; Available from: 10.1016/S2542-5196(21)00350-8
17. Andreão WL, Toledo de Almeida Albuquerque T. Avoidable mortality by implementing more restrictive fine particles standards in Brazil: An estimation using satellite surface data. Environ Res. 2021 Jan;192:110288. [PubMed: 33038364]
18. Zhang H, Kondragunta S. Daily and Hourly Surface PM2.5 Estimation From Satellite AOD. Earth and Space Science. 2021;8(3):e2020EA001599.
19. Diao M, Holloway T, Choi S, O'Neill SM, Al-Hamdan MZ, Van Donkelaar A, et al. Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models. Journal of the Air & Waste Management Association. 2019;69(12):1391–414. [PubMed: 31526242]

20. Gutiérrez-Avila I, Arfer KB, Wong S, Rush J, Kloog I, Just AC. A spatiotemporal reconstruction of daily ambient temperature using satellite data in the Megalopolis of Central Mexico from 2003 to 2019. Int J Climatol. 2021 Jun 30;41(8):4095–111. [PubMed: 34248276]
21. Jáuregui E The Climate of the Mexico City Air Basin: Its Effects on the Formation and Transport of Pollutants. In: Fenn ME, de Bauer LI, Hernández-Tejeda T, editors. Urban Air Pollution and Forests: Resources at Risk in the Mexico City Air Basin. New York, NY: Springer New York; 2002. p. 86–117.
22. Lyapustin A, Wang Y. MCD19A2 MODIS/Terra+aqua land aerosol optical depth daily L2G global 1km SIN grid V006 [Internet]. NASA EOSDIS Land Processes DAAC; 2018. Available from: https://lpdaac.usgs.gov/products/mcd19a2v006/
23. Thornton MM, Shrestha R, Wei Y, Thornton PE, Kao SC, Wilson BE. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 [Internet]. ORNL DAAC; 2020 [cited 2021 Oct 29]. Available from: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1840
24. Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, et al. ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [Internet]. [cited 2021 Oct 29]. Available from: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview
25. OpenStreetMap Wiki contributors. Main page [Internet]. OpenStreetMap Wiki; [cited 2021 Oct 29]. Available from: https://wiki.openstreetmap.org/w/index.php?title=Main_Page&oldid=1060762
26. Chen T, Guestrin C. XGBoost [Internet]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. Available from: 10.1145/2939672.2939785
27. Carnell R Latin Hypercube Samples [R package lhs version 1.1.3]. 2021 [cited 2022 Jun 10]; Available from: https://CRAN.R-project.org/package=lhs
28. Baston D exactextractr: Fast extraction from raster datasets using polygons [Internet]. 2020. Available from: https://cran.r-project.org/web/packages/exactextractr/exactextractr.pdf
29. Center for International Earth Science Information Network-CIESIN-Columbia University. Gridded population of the world, version 4 (GPWv4): population density [Internet]. 2016. Available from: https://sedac.ciesin.columbia.edu/data/collection/gpw-v4/documentation
30. Secretaría de Salud. Norma Oficial Mexicana NOM-025-SSA1-2021, Salud ambiental. Valores límite permisibles para la concentración de partículas suspendidas PM10 y PM2.5 en el aire ambiente y criterios para su evaluatión [Internet]. 2021 Oct. Available from: https://dof.gob.mx/nota_detalle.php?codigo=5633855&fecha=27/10/2021
31. Consejo Nacional de Población. Datos Abiertos del Índice de Marginación [Internet]. 2013 [cited 2021 Oct 27]. Available from: http://www.conapo.gob.mx/es/CONAPO/Datos_Abiertos_del_Indice_de_Marginacion
32. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. 2022. Available from: https://www.R-project.org
33. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 0 4-2. 2015;1(4):1–4.
34. USAID. Actualización al diagnóstico de la Megalópolis del centro de México. Mexico Low Emissions Development Program (MLED) [Internet]. 2014. Available from: http://www.plataformaleds.org/productos-programa-mled.php
35. Trejo-González AG, Riojas-Rodriguez H, Texcalac-Sangrador JL, CM Guerrero-López, Cervantes-Martínez K, Hurtado-Díaz M, et al. Quantifying health impacts and economic costs of PM2.5 exposure in Mexican cities of the National Urban System. Int J Public Health. 2019 May;64(4):561–72. [PubMed: 30834460]
36. Office of Air Quality Planning and Standards Health and Environmental Impacts Division Research Triangle Park, NC. Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter [Internet]. U.S. EPA. ; 2020 Jan. Report No.: EPA-452/R-20-002. Available from: https://www.epa.gov/system/files/documents/2021-10/final-policy-assessment-for-the-review-of-the-pm-naaqs-01-2020.pdf

37. Xiao Q, Chang HH, Geng G, Liu Y. An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data. Environ Sci Technol. 2018 Nov 20;52(22):13260–9. [PubMed: 30354085]
38. Nielsen D Tree boosting with xgboost-why does xgboost win "every" machine learning competition? [Internet]. ntnuopen.ntnu.no; 2016 [cited 2022 Jun 7]. Available from: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf
39. SEMARNAT, SEDEMA, SMAGEM, SEMARNATH. Programa de Gestión para Mejorar la Calidad del Aire de la Zona Metropolitana del Valle de México (ProAire ZMVM 2021- 2030) [Internet]. 2021 Dec [cited 2022 May 4]. Available from: https://dsiappsdev.semarnat.gob.mx/datos/portal/proaire/2022/38_ProAire_ZMVM.pdf
40. Secretaría del Medio Ambiente. Historical Analysis of Population Health Benefits Associated with Air Quality in Mexico City during 1990 and 2015 [Internet]. 2018. Available from: http://www.data.sedema.cdmx.gob.mx/beneficios-en-salud-por-la-mejora-de-la-calidad-del-aire/descargas/analisis-ingles.pdf
41. Martenies SE, Wilkins D, Batterman SA. Health impact metrics for air pollution management strategies. Environ Int. 2015 Dec;85:84–95. [PubMed: 26372694]
42. Molina LT, Velasco E, Retama A, Zavala M. Experience from Integrated Air Quality Management in the Mexico City Metropolitan Area and Singapore. Atmosphere . 2019 Aug 31;10(9):512.
43. Zalakeviciute R, López-Villada J, Rybarczyk Y. Contrasted Effects of Relative Humidity and Precipitation on Urban PM2.5 Pollution in High Elevation Urban Areas. Sustain Sci Pract Policy. 2018 Jun 18;10(6):2064.
44. Kinney PL. Interactions of Climate Change, Air Pollution, and Human Health. Curr Environ Health Rep. 2018 Mar;5(1):179–86. [PubMed: 29417451]
45. Yitshak-Sade M, Bobb JF, Schwartz JD, Kloog I, Zanobetti A. The association between short and long-term exposure to PM2.5 and temperature and hospital admissions in New England and the synergistic effect of the short-term exposures. Sci Total Environ. 2018 Oct 15;639:868–75. [PubMed: 29929325]
46. Kioumourtzoglou MA, Schwartz J, James P, Dominici F, Zanobetti A. PM2.5 and Mortality in 207 US Cities: Modification by Temperature and City Characteristics. Epidemiology. 2016 Mar;27(2):221–7. [PubMed: 26600257]
47. Godwin W Assessment of interactive effects of temperature and air pollution on mortality in Mexico City [Internet] [Master of Public Health]. Stanaway J, editor. University of Washington ; 2018. Available from: https://digital.lib.washington.edu/researchworks/handle/1773/42894
48. World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary. 2021.
49. Son JY, Bell ML. The relationships between short-term exposure to particulate matter and mortality in Korea: Impact of particulate matter exposure metrics for sub-daily exposures. Environ Res Lett. 2013 Mar;8(1):014015. [PubMed: 25580157]
50. Lin H, Liu T, Xiao J, Zeng W, Guo L, Li X, et al. Hourly peak PM2.5 concentration associated with increased cardiovascular mortality in Guangzhou, China. J Expo Sci Environ Epidemiol. 2017 May;27(3):333–8. [PubMed: 27805624]
51. Keller JP, Chang HH, Strickland MJ, Szpiro AA. Measurement Error Correction for Predicted Spatiotemporal Air Pollution Exposures. Epidemiology. 2017 May;28(3):338–45. [PubMed: 28099267]
52. Zoogman P, Liu X, Suleiman RM, Pennington WF, Flittner DE, Al-Saadi JA, et al. Tropospheric emissions: Monitoring of pollution (TEMPO). J Quant Spectrosc Radiat Transf. 2017 Jan 1;186:17–39. [PubMed: 32817995]
53. Naeger AR, Newchurch MJ, Moore T, Chance K, Liu X, Alexander S, et al. Revolutionary Air-Pollution Applications from Future Tropospheric Emissions: Monitoring of Pollution (TEMPO) Observations. Bull Am Meteorol Soc. 2021 Sep 1;102(9):E1735–41.

Significance.

Machine learning algorithms can be used to predict highly spatiotemporally resolved $PM_{2.5}$ concentrations even in regions with sparse monitoring.

Impact.

Our $PM_{2.5}$ predictions can aid local authorities in air-quality management and public-health surveillance, and they can advance epidemiological research in Central Mexico with state-of-the-art exposure assessment methods.

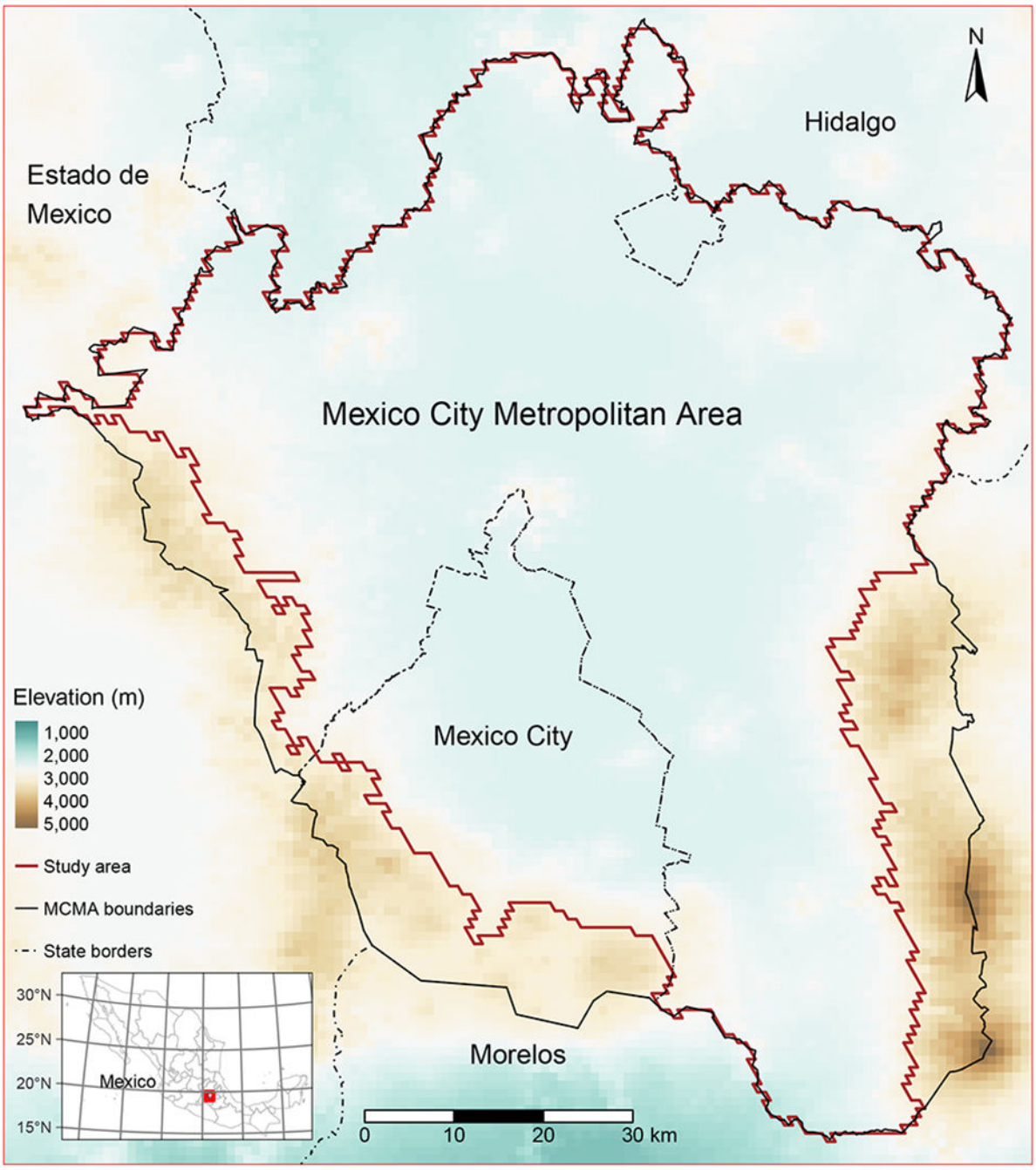

Figure 1.
Map of study area in Central Mexico. The study area used for our $PM_{2.5}$ models in the Mexico City Metropolitan Area (MCMA).

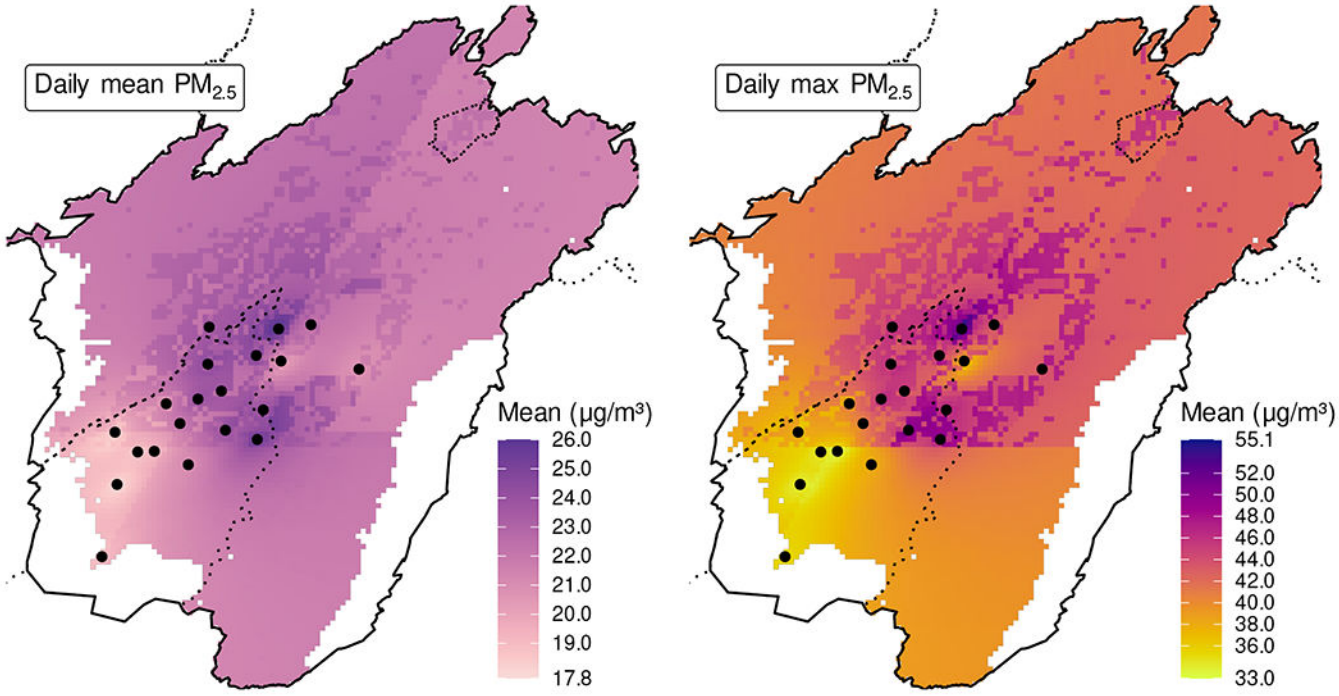

Figure 2.

Maps of the averaged annual daily mean and daily max $PM_{2.5}$ concentrations for 2019 in the Mexico City Metropolitan Area. Solid and dotted lines indicate the Mexico City Metropolitan Area and Mexican states boundaries, respectively. Black dots indicate ground monitors.

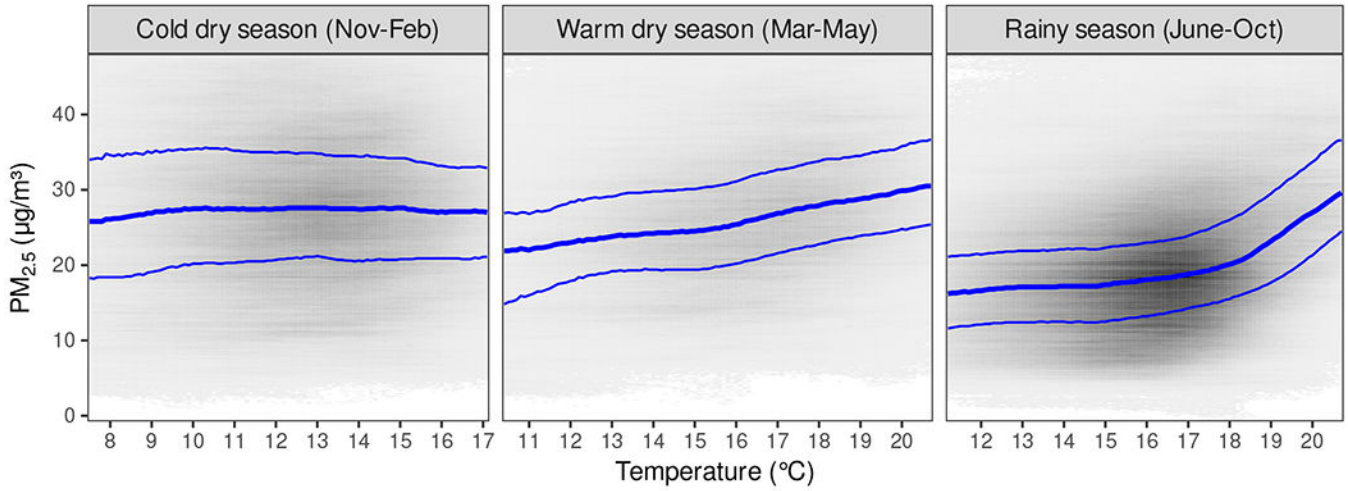

Figure 3.

Heatmaps of mean temperature and mean $PM_{2.5}$, counting all grid cells and days equally. Darker areas indicate more grid cells, more days, or both. Temperature and $PM_{2.5}$ predictions are already rounded to the nearest tenth, so no further grouping is needed for a heatmap. For legibility, the temperature scale only shows the middle 95% of the data for each season, and the $PM_{2.5}$ scale only goes up to the 98^{th} percentile for all seasons. Blue lines show the quartiles of $PM_{2.5}$ conditional on temperature.

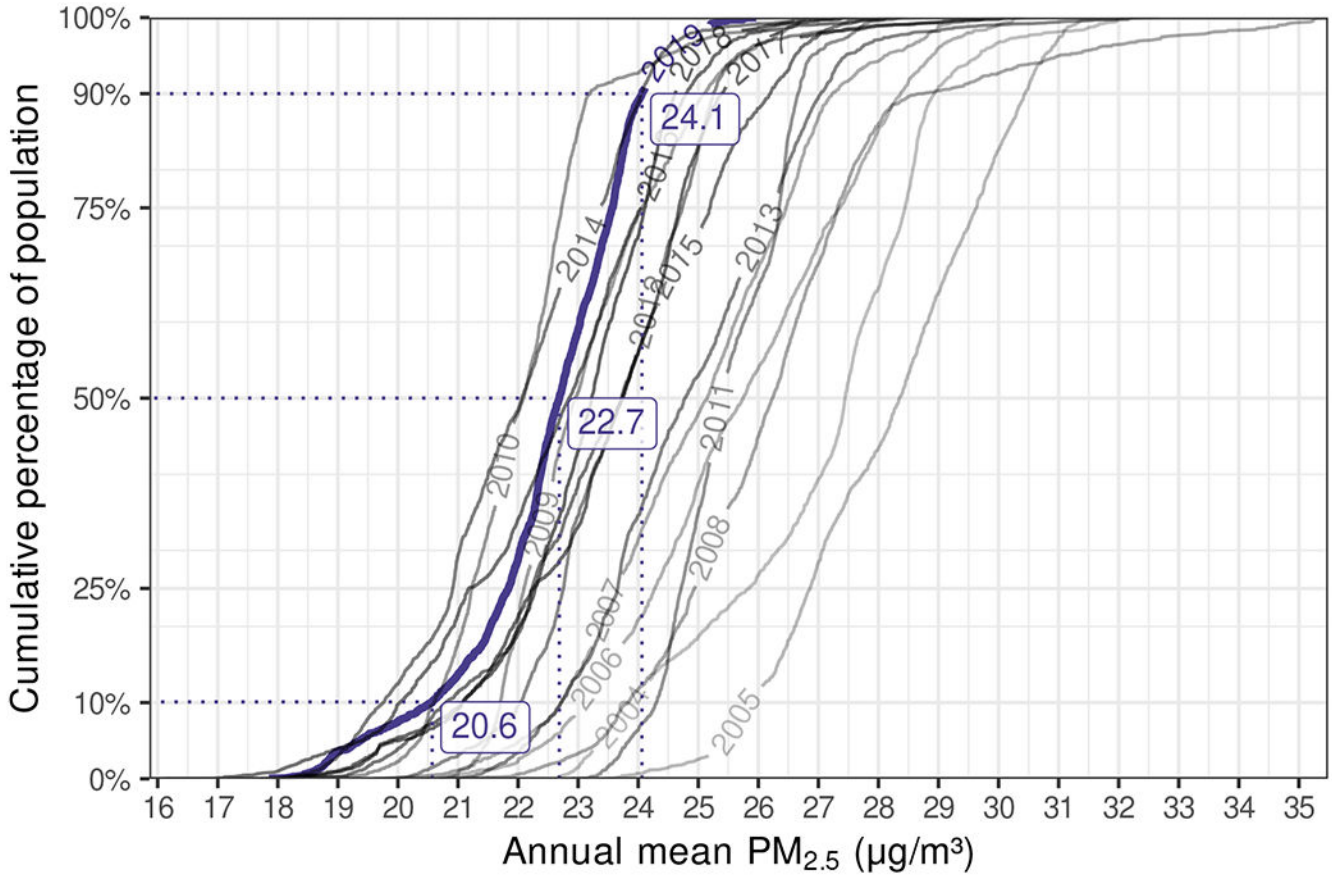

Figure 4.

Population estimated annual average exposures. The figure shows an empirical cumulative distribution curve for each year from 2004 to 2019, generated from our daily mean model and using the 2010 census population density. Specific quantiles are labeled for the year 2019, where only 10% of the population in the study region had an annual average exposure below 20.6 µg/m^3.

Table 1.

Assessment of cross-validated predictions from the daily mean $PM_{2.5}$ model by year

Year	Number of stations	Observations	R^2	SD	RMSE	MAD	MAE
2004	8	2,751	0.76	12.02	5.86	9.12	3.91
2005	8	2,701	0.81	14.80	6.43	11.28	4.38
2006	8	2,685	0.68	13.19	7.48	9.55	5.04
2007	9	2,855	0.71	10.87	5.85	8.16	4.28
2008	9	3,040	0.64	12.16	7.29	9.27	4.61
2009	9	2,670	0.75	10.14	5.09	7.71	3.61
2010	9	2,844	0.79	11.70	5.41	8.83	3.64
2011	12	3,019	0.77	11.53	5.56	8.90	3.88
2012	13	4,025	0.76	10.10	4.95	7.63	3.59
2013	13	4,362	0.80	11.75	5.25	8.85	3.87
2014	14	4,203	0.73	9.87	5.10	7.50	3.86
2015	19	5,194	0.77	10.78	5.11	7.90	3.76
2016	17	5,307	0.83	11.44	4.73	8.56	3.37
2017	17	4,901	0.80	10.79	4.78	8.42	3.15
2018	17	4,633	0.84	9.91	3.94	7.19	2.83
2019	20	5,175	0.86	11.50	4.26	7.98	2.85

Standard deviation (SD), Root mean squared error (RMSE), Mean absolute deviation (MAD), and Mean Absolute Error (MAE)

J Expo Sci Environ Epidemiol. Author manuscript; available in PMC 2023 March 10.

Table 2.

Assessment of cross-validated predictions from the daily one-hour maximum $PM_{2.5}$ model by year

Year	Number of stations	Observations	R^2	SD	RMSE	MAD	MAE
2011	12	3,019	0.47	24.26	17.63	16.65	10.36
2012	13	4,025	0.46	21.80	16.09	15.18	10.17
2013	13	4,362	0.58	23.78	15.49	17.28	10.27
2014	14	4,203	0.52	19.54	13.58	14.46	9.76
2015	19	5,194	0.63	25.30	15.34	16.37	9.97
2016	17	5,307	0.62	25.33	15.59	16.48	8.68
2017	17	4,901	0.56	23.86	15.75	15.85	8.48
2018	17	4,633	0.63	19.68	11.96	13.74	7.83
2019	20	5,175	0.66	21.39	12.50	14.08	8.04

Standard deviation (SD), Root mean squared error (RMSE), Mean absolute deviation (MAD), and Mean Absolute Error (MAE)