





















# Nationwide estimation of daily ambient PM<sub>2.5</sub> from 2008 to 2020 at 1 km<sup>2</sup> in India using an ensemble approach

Siddhartha Mandal <sup>a,b,1</sup>, Ajit Rajiva <sup>b,1</sup>, Itai Kloog <sup>c</sup>, Jyothi S. Menon <sup>b</sup>, Kevin J. Lane <sup>d</sup>, Heresh Amini <sup>e</sup>, Gagandeep K. Walia <sup>a,b</sup>, Shweta Dixit <sup>b</sup>, Amruta Nori-Sarma <sup>d</sup>, Anubрати Dutta <sup>a,b</sup>, Praggya Sharma <sup>a</sup>, Suganthi Jaganathan <sup>a,b,f</sup>, Kishore K. Madhipatla <sup>g</sup>, Gregory A. Wellenius <sup>d</sup>, Jeroen de Bont <sup>f</sup>, Chandra Venkataraman <sup>h</sup>, Dorairaj Prabhakaran <sup>a,b</sup>, Poornima Prabhakaran <sup>a,b,2</sup>, Petter Ljungman <sup>f,i,\*2</sup> and Joel Schwartz <sup>j,2</sup>

<sup>a</sup>Centre for Chronic Disease Control, New Delhi 110016, India

<sup>b</sup>Public Health Foundation of India, New Delhi 110017, India

<sup>c</sup>Department of Environmental, Geoinformatics and Urban Planning Sciences, Ben Gurion University of the Negev, Beer Sheva 84105, Israel

<sup>d</sup>Department of Environmental Health, Boston University School of Public Health, Boston, MA 02118, USA

<sup>e</sup>Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

<sup>f</sup>Institute of Environmental Medicine, Karolinska Institute, Stockholm 17177, Sweden

<sup>g</sup>Center for Atmospheric Particle Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

<sup>h</sup>Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India

<sup>i</sup>Department of Cardiology, Danderyd Hospital, Stockholm 18257, Sweden

<sup>j</sup>Department of Environmental Health, Harvard TH Chan School of Public Health, Harvard University, Boston, MA 02115, USA

\*To whom correspondence should be addressed: Email: [petter.ljungman@ki.se](mailto:petter.ljungman@ki.se)

<sup>1</sup>S.M. and A.R. are joint first authors.

<sup>2</sup>P.P., P.L., and J.S. are joint last author.

**Edited By:** Doraiswami Ramkrishna

## Abstract

High-resolution assessment of historical levels is essential for assessing the health effects of ambient air pollution in the large Indian population. The diversity of geography, weather patterns, and progressive urbanization, combined with a sparse ground monitoring network makes it challenging to accurately capture the spatiotemporal patterns of ambient fine particulate matter (PM<sub>2.5</sub>) pollution in India. We developed a model for daily average ambient PM<sub>2.5</sub> between 2008 and 2020 based on monitoring data, meteorology, land use, satellite observations, and emissions inventories. Daily average predictions at each 1 km × 1 km grid from each learner were ensemble using a Gaussian process regression with anisotropic smoothing over spatial coordinates, and regression calibration was used to account for exposure error. Cross-validating by leaving monitors out, the ensemble model had an R<sup>2</sup> of 0.86 at the daily level in the validation data and outperformed each component learner (by 5–18%). Annual average levels in different zones ranged between 39.7 μg/m<sup>3</sup> (interquartile range: 29.8–46.8) in 2008 and 30.4 μg/m<sup>3</sup> (interquartile range: 22.7–37.2) in 2020, with a cross-validated (CV)-R<sup>2</sup> of 0.94 at the annual level. Overall mean absolute daily errors (MAE) across the 13 years were between 14.4 and 25.4 μg/m<sup>3</sup>. We obtained high spatial accuracy with spatial R<sup>2</sup> greater than 90% and spatial MAE ranging between 7.3–16.5 μg/m<sup>3</sup> with relatively better performance in urban areas at low and moderate elevation. We have developed an important validated resource for studying PM<sub>2.5</sub> at a very fine spatiotemporal resolution, which allows us to study the health effects of PM<sub>2.5</sub> across India and to identify areas with exceedingly high levels.

**Keywords:** India, particulate matter, high resolution, spatiotemporal, machine learning

## Significance Statement

High levels of particulate matter (PM<sub>2.5</sub>) are a major public health hazard in a populous country like India. However, sparse ground monitoring and lack of detailed exposure assessments present major hurdles to understand the health effects of PM<sub>2.5</sub>. In this paper, we have developed a model for assessing daily ambient PM<sub>2.5</sub> at 1 km × 1 km across India from 2008 to 2020, with high accuracy. We used PM<sub>2.5</sub> data from monitoring stations, predictors from multiple domains, along multiple machine-learning algorithms to predict PM<sub>2.5</sub> levels at high spatiotemporal resolution, while ensuring the representativeness of the model. This presents a valuable resource to generate evidence on health effects of PM<sub>2.5</sub> across urban, periurban, and rural India, which is critical for informing policy actions.

**Competing Interest:** The authors declare no competing interest.

**Received:** August 15, 2023. **Accepted:** February 16, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

## Introduction

Fine particulate matter (PM<sub>2.5</sub>) has been linked with multiple health outcomes across the life course, including those related to pregnancy (1), birth (2), cardiovascular (3), cardiometabolic (4), pulmonary (5), and cognitive conditions (6). Mechanistic studies indicate that PM<sub>2.5</sub> affects health through multiple pathways, such as inflammation, oxidative stress, imbalance of the autonomic nervous system, direct translocation, and DNA methylation (7, 8). High levels of PM<sub>2.5</sub> and other air pollutants have been a major public health problem in India (9), with concentrations exceeding more than 10 times the World Health Organization (WHO) recommended levels in certain areas (10). Further, high levels of PM<sub>2.5</sub> are not limited to urban areas and affect a large share of the population across the urban–rural gradient (11). However, most of the research linking PM<sub>2.5</sub> with health has been conducted in countries with lower levels of PM<sub>2.5</sub>, with relatively less research originating from low- and middle-income countries. Differences in particle composition and nonlinearities observed in concentration–response relationships make extrapolating the results from higher income and lower concentration countries difficult.

A major obstacle in conducting research on health effects of air pollution in India is the lack of robust exposure estimates at fine spatiotemporal resolution across the country. Existing estimates of PM<sub>2.5</sub> levels across India are inadequate due to the low volume of data from a sparse monitoring network (12) and limitations of existing methods for predicting local levels, such as land-use regression, emissions inventories, and chemical transport models (CTM) (13). Important recent advances include a satellite-informed machine-learned model for the state of Delhi over 7 years (14), a model for the Indo-Gangetic Plain using data from a single year and based mostly on meteorological predictors (15), and models based on aerosol optical depth (AOD) over India (16). Several of these models are reliant on global CTM (11, 17) and use a calibration approach based on the ratio between PM<sub>2.5</sub> and AOD. Despite these recent advancements leveraging remote sensing, there are limitations that need to be addressed, with respect to spatiotemporal resolution and methodologies. Specifically, models relying mainly on satellite observations (such as AOD), sparse ground monitoring data, or a few meteorological parameters are inadequate to fully capture the localized spatiotemporal trends in PM<sub>2.5</sub> across a geographically diverse country like India, which spans more than 3 million km<sup>2</sup>. Further, predictions using only the coincident ratio between CTM-predicted PM<sub>2.5</sub> and AOD inherently assume that AOD is an accurate predictor of PM<sub>2.5</sub>, while ignoring variables such as road networks, vegetation, and fires, which have been shown as major contributors to PM<sub>2.5</sub> concentrations. Further, the use of global CTM often does not incorporate local sources of air pollution and relies on linear regression which is known to have limitations in modeling complex relationships. These approaches also assume that the effect of meteorology on PM<sub>2.5</sub> is entirely captured through the CTM.

Another approach to model PM<sub>2.5</sub> uses a hierarchical Bayesian model to calibrate annual average PM<sub>2.5</sub> concentrations using a global dataset and chemical transport model (18). Modeling annual averages rather than 24-h averages ignores the short-term variations in PM<sub>2.5</sub>. With several country and city-specific models showing significant variations in sources and patterns of air pollution, global models often fail to capture finer local variations that are important for individual-level epidemiological studies on acute and chronic exposure to air pollution (19, 20). Studies on pregnancy and birth outcomes (1), neurodevelopment (21),

hypertension (22), and diabetes (23) often need different durations of exposure including short- and long-term averages to understand the associations. Further, the high uncertainty from existing global models in South Asia (18) makes it important to have country-specific models that incorporate more ground monitoring data as well as region-specific variables within methods that are adept at modeling complex relationships to arrive at the PM<sub>2.5</sub> estimates.

Moreover, to the extent that the estimates from PM<sub>2.5</sub> prediction models are to be used in epidemiologic health effects studies, it is important to ensure that the resulting exposure estimates will have minimum bias and relative error, and generate minimal bias when used in epidemiology studies. Although methods such as regression calibration (24) are available to adjust for exposure measurement error in epidemiology studies, exposure assessment models seldom implement any such corrections, thus potentially leading to bias in the epidemiological associations reported in health effect studies utilizing the predictions.

To address these gaps in exposure assessment in India, here we describe the development of a novel nationwide model to estimate daily ambient PM<sub>2.5</sub> concentrations at a 1 km × 1 km spatial resolution between 2008 and 2020. Our approach leverages an unprecedented set of features (meteorological, built environment, remote-sensed, and chemical transport model outputs) as inputs into a series of machine-learned models which are calibrated against a large curated database of daily average PM<sub>2.5</sub> data from ground-based monitors across India. Regression calibration applied to the ensemble average from these models provides highly localized, minimally biased estimates of daily PM<sub>2.5</sub> for nearly every square kilometer in India. This novel dataset is already being used to enable previously infeasible research, increase local awareness of air pollution levels across India, and influence policy.

## Methods

Data on daily average PM<sub>2.5</sub> and PM<sub>10</sub> were collected from the air quality monitoring stations maintained by the Central and State Pollution Control Boards of India; real-time monitors at embassy of the United States of America in New Delhi, Mumbai, Hyderabad, Chennai, and Kolkata; and data collected as part of academic campaigns (Details in [Supplemental Material](#)). Data from both continuous real-time monitoring stations as well as manual monitoring stations were included in the analysis. To address the sparseness in observed daily PM<sub>2.5</sub>, we first developed an imputation model based on an extreme gradient boosting algorithm (25), for the ratio of daily average PM<sub>2.5</sub> and PM<sub>10</sub> at locations and days where measurements of both fractions were available. The details of the predictors used in this model are provided in the [Supplemental Material](#). This calibration model was trained on 172,983 available ratios and had an overall cross-validated (CV) R<sup>2</sup> of 0.91 in a left-out validation dataset of 43,245 observations. This was used to predict PM<sub>2.5</sub> at locations and times where only PM<sub>10</sub> was recorded, thus extending monitor coverage across India (additional 387,883 observations of PM<sub>2.5</sub>). Details of quality checking for the monitored data are provided in the [Supplemental Material](#).

To account for zonal similarities and differences, the country was classified into 6 zones similar to the boundaries defined by the State Reorganisation Act 1956, Government of India, Section 15, with the addition of the North-East Zone as per the North Eastern Council Act. The notable exception here was that of Sikkim being added to the Eastern Zone for spatial consistency.

The resulting zone variable was used in the model as a categorical predictor. For model summaries, we also used districts which are second-level administrative divisions within the 29 states of the country.

We incorporated predictors across multiple domains including satellite-based observations, meteorology, land-use patterns, and emissions inventories. Table 1 shows the major domains, sources, and spatial and temporal resolution of the predictors. All variables at coarser spatial resolution were downscaled to the 1 km × 1 km grids of interest before developing the model. Variables available at resolutions finer than 1 km × 1 km were aggregated to ensure all variables were available at the same spatial resolution. Further details describing the predictors are provided in the [Supplemental Material](#). The model development was a multi-stage process described in detail below and a schematic diagram in Fig. 1 demonstrates the same process.

### Aerosol optical depth imputation

The daily multiangle implementation of atmospheric correction (MAIAC) AOD (470 nm) at 1 km × 1 km resolution was retrieved from the moderate resolution imaging spectroradiometer (MODIS) instrument of the Terra and Aqua satellites along with quality assurance flags from NASA's Land Processes Distributed Active Archive Center (LP DAAC) at the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center (<https://lpdaac.usgs.gov/products/mcd19a2v061/>). Preprocessing was carried out to remove observations contaminated by cloud and snow (removing cloudy pixels and observations surrounded by more than >8 cloudy pixels) and filtering outliers based on the valid MAIAC AOD values (removing AOD values <0 and > 1.2), as done in several similar exposure studies to avoid pixels contaminated by clouds or snow (36, 37). After these corrections, it was observed that more than 60% of AOD data were missing over the study region for the study period. To impute the missing AOD, the MODIS AOD observations were calibrated against the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis-based total AOD at 469 nm (at a spatial resolution of 80 km) using a deep-learning algorithm

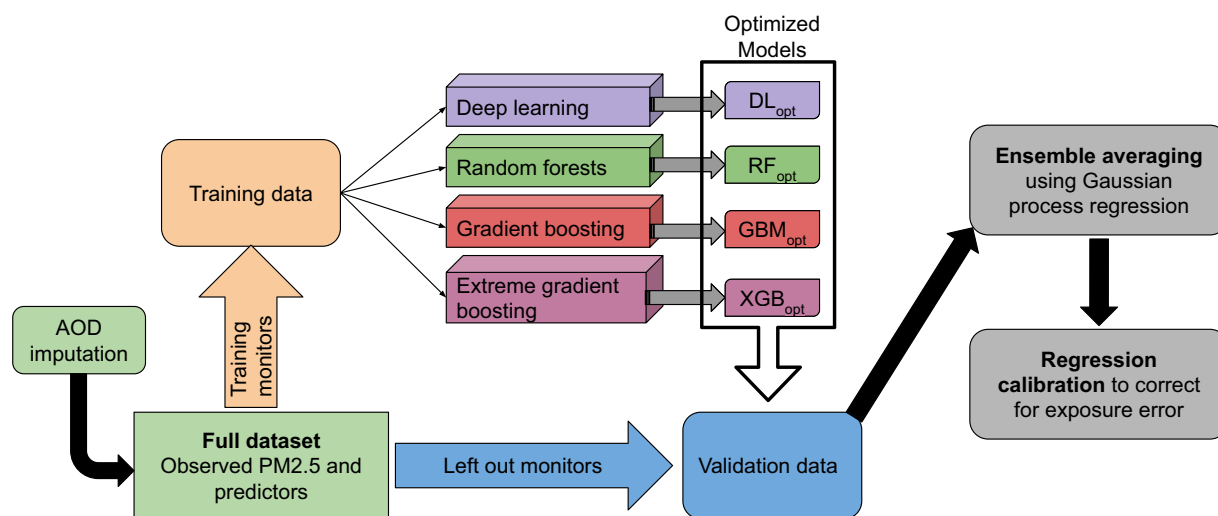
(38, 39), along with other predictors, such as meteorological variables, elevation, and population density. Specifically, the deep-learning model was a multilayer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. We used the AOD at 469 nm since several publications have shown that the deep blue algorithm (which utilized the blue wavelengths, e.g. 470 nm) for MODIS performed better than the original over bright surfaces (e.g. urban areas and deserts) (36). The best-performing model was used to predict the AOD values for missing days and grid cells. Annual CV R<sup>2</sup> for this model was in the range of 0.73–0.87 across 2008–2020 while the same during monsoon (June–September) was in the range of 0.64–0.81 across 2008–2020 (Table S1). In the second stage of modeling the relationship between PM<sub>2.5</sub> and the predictors, we used the MAIAC AOD wherever available along with the imputed AOD when MAIAC AOD was missing. When using this imputed AOD in our machine learning models, we included imputed AOD up to 2 to allow for potential high-pollution scenarios. That is, we believe a measured AOD of 2 is too likely to represent cloud contamination rather than high air pollution, but an imputed AOD of 2 based on the variables we used is not subject to cloud contamination and likely represents a very high air pollution day.

### Cross-validation procedure

We split the entire dataset by sampling monitors which provided observed PM<sub>2.5</sub> (or PM<sub>2.5</sub> from the PM ratio model). Given the concentration of stationary monitors in a few urban locations (for example, Delhi), especially in the earlier years (2008–2016), we needed to ensure that our training set did not oversample monitors with large amounts of data concentrated in the North Zone. This inequality in the spatial distribution of the monitors across India is the reason why we did not sample equally across zones. We created quintile-based strata of data availability for each station (ratio of monitor-specific observations and total observations). Within each stratum, we sampled 80% of the available monitors while leaving out 20% of the monitors as validation dataset. This resulted in 211 stations (out of 1,056) being included in

**Table 1.** Variables used during model development are segregated by major domains along with corresponding sources, spatial, and temporal resolution.

Type	Variable	Resolution		Source
		Spatial	Temporal	
Satellite-based	AOD	0.01° × 0.01°	Daily	MAIAC products from MODIS (26)
	Vegetation index (NDVI)	0.01° × 0.01°	16-day	MODIS
	Active Fires	0.01° × 0.01°	Daily	MODIS Active Fires (27)
	Light intensity at night	0.01° × 0.01°	Annual	Visible Infrared Imaging Radiometer Suite (VIIRS) (28)
	NO <sub>2</sub> concentrations	0.01° × 0.01°	2019 annual average	Sentinel 5P (29)
Meteorology	Reanalysis-based variables (ECMWF Re-Analysis [ERA] 5)	0.125° × 0.125°	Daily	European Centre for Medium Range Weather Forecast (ECMWF) (30)
Land use	Road density	1 km × 1 km	Time-invariant	Open Street Maps
	Population density	1 km × 1 km	2010, 2015, 2020	Gridded Population of the World version 4 (31)
	Elevation	30 m × 30 m	Time-invariant	Terra Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) Version 3 (32)
Emissions	Location of toll plazas, airports, power plants	—	Time-invariant	
	Land classification	Categorical	Time-invariant	Census based
	Sectorwise estimates of PM <sub>2.5</sub> emissions (in tons/year)	0.25° × 0.25°	5 year average	Venkataraman et al. (33)
Reanalysis-based variables	Carbon emissions from fires	0.25° × 0.25°	Daily	GFED (34)
	AOD and PM <sub>2.5</sub>	80 km × 80 km	Daily	CAMS (35)



**Fig. 1.** Model development framework: Machine-learning-based framework for developing the spatiotemporal prediction model for daily ambient  $PM_{2.5}$ , starting with AOD imputation, splitting of data by monitors, training of the four learners, application of optimized models on validation data, ensemble averaging, and regression calibration to correct for exposure error.

the left-out validation dataset, the rest was the training dataset. The zonal distribution of the stations is shown in Table S2.

## Developing the learners

The relationship between observed  $PM_{2.5}$  and all predictors in the training data was modeled using machine-learning algorithms, including deep-learning (39), random forests (40), gradient boosting (41), and extreme gradient boosting (25). Within each algorithm, an internal cross-validation (by splitting the training dataset using a 90:10 split) was implemented to ensure optimal selection of the hyperparameters and prevent overfitting. The choice of hyperparameter space for each algorithm is provided in the [Supplemental Material](#). The best learner within each algorithm was obtained by minimizing root-mean-squared errors. Using each optimized learner, predictions were obtained on the left-out validation data. Predictions from the four machine learners were combined using an ensemble model based on Gaussian processes (allowing for anisotropic smoothing across the coordinates) which utilized the predictions from each learner, month, year, elevation, vegetation, and density of roads in a 10 km buffer as predictors (42, 43). We obtained  $R^2$  (overall, spatial, and temporal), mean absolute errors (overall, spatial, and temporal mean absolute error [MAE]), slope, and bias of the predictions within the test dataset to assess prediction accuracy of each component learner and the ensemble-averaged model.

## Analyzing variation (error) of predictions

We computed monthly average standard deviations of the residuals between predicted and observed  $PM_{2.5}$  and analyzed these variations against meteorological and land-use variables to understand the factors associated with the performance of the model. Generalized linear models were used with a gamma distribution because of the positively skewed nature of the outcome while accounting for overdispersion.

## Correcting for exposure error

A regression calibration approach was implemented to account for exposure error in the predictions (44). For this, the grid cells were classified into clusters based on the similarity of land-use characteristics (elevation, land classification, census classification, and

distance to major and medium cities) using a clustering for mixed variable-type data (k-proto clustering) (45). In the left-out validation dataset, a robust linear regression was used to calibrate the ground monitoring data with the ensemble-averaged predictions, stratified by cluster and month. The optimum number of clusters was selected by minimizing the within-cluster sum of squares. The slopes from these models were used to upscale or downscale the ensemble-averaged predictions to provide concentrations corrected for bias and exposure error. Details of the method are provided in the [Supplementary materials](#).

## Results

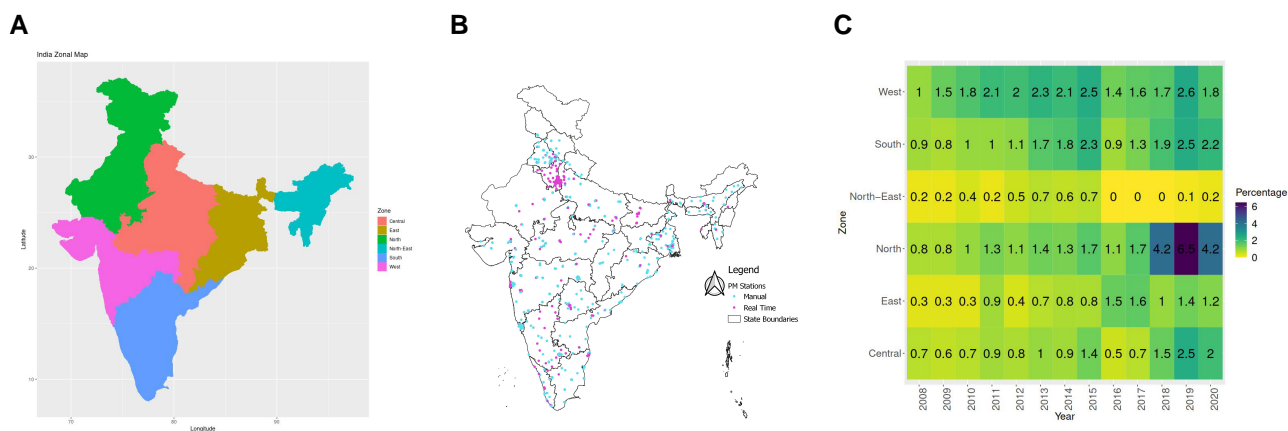
### Description of the ground monitoring data

Figure 2 describes the ground monitoring data for  $PM_{2.5}$ , which amounted to 679,354 observations from 1,056 stations (both continuous real-time and manual). The locations of the monitors are shown in Fig. 1B. Panel (C) shows the availability of data by Zone and year, which indicates that 15% of the total ground monitoring data were from 2018 to 2020 in the North Zone. We also analyzed the growth in the number of continuous real-time monitoring stations, which are installed and maintained by the Central Pollution Control Board, across time and observed a steep increase starting from 2015. Initially, continuous monitors were only present in two districts in India (Delhi and Mumbai) with 0.6 (~1) monitors being added each year to unmonitored districts. Since 2015, ~21 monitors have been added each year up to 2020 (Fig. S2).

### Model performance

We evaluated the model performances for the individual learner models as well as the regression-calibrated ensemble-averaged models within the left-out validation dataset by computing metrics shown in Table 2A. The overall  $CV-R^2$  for daily predictions for the entire country over the 13-year period was 0.86, with performance improving across the years. Splitting the performance by zone, we observed that the ensemble-averaged model performed best in the North (0.87) and Central (0.83) zones compared to others. We note that among the learners, gradient boosting-based models out-performed deep-learning and random forests, across all years. Importantly, the ensemble-averaged model had a higher daily  $CV-R^2$  than each component learner during the





**Fig. 2.** PM<sub>2.5</sub> and PM<sub>10</sub> ground monitoring stations and data availability: A) Zonal breakup of India into six zones, North, Central, East, North-East, South, and West. B) Location of manual and real-time ground monitoring stations that provided daily average PM<sub>2.5</sub> and PM<sub>10</sub> observations in India. C) Proportion of total number of PM<sub>2.5</sub> observations from ground monitoring data according to the six zones and 13 years.

entire period. The improvement in daily CV-R<sup>2</sup> from using the ensemble average was 18%, 7%, 7.5%, and 5% in comparison to deep-learning, random forests, extreme gradient boosting, and gradient boosting, respectively. The models performed better in terms of spatial accuracy with spatial R<sup>2</sup> > 0.90 across all years, while temporal accuracy was lower in the earlier years. The overall daily MAE across the years ranged from 14.4 to 25.4 µg/m<sup>3</sup>, while the spatial MAEs were observed to be 7.3–16.5 µg/m<sup>3</sup> (Table 2B). Bias in the predictions were between 5.3 and 10.6 µg/m<sup>3</sup> across years, while all the slopes of the regression between observed and predicted daily PM<sub>2.5</sub> concentrations were close to 1 (Table 2C). For annual average predictions, the CV-R<sup>2</sup> and MAE were 0.94 and 8.8 µg/m<sup>3</sup>, respectively. Using a regression calibration approach by spatial clusters and month, we obtained coefficients (shown in Table S1) to upscale and downscale the ensemble-averaged predictions to provide predictions corrected for exposure error.

### Analysis of precision and regression calibration

We analyzed the monthly standard deviation of the predictions at each station in the validation dataset against spatial predictors (Fig. S3). Using a nonlinear regression model, we observed precision was lower at rural and vegetation areas compared to urban areas, at grid cells with high elevation, and those away from large and medium cities. Among nonlinear associations, precision was lower in areas with high vegetation and median NO<sub>2</sub> concentrations (from Sentinel 5P). We also assessed the prediction accuracy of the model between urban and rural stations in the validation data (Table S3), where we observed the accuracy was higher at stations located in statutory towns (daily CV-R<sup>2</sup> > 0.77 across all years), whereas in rural stations the accuracy was lower (daily CV-R<sup>2</sup> < 0.65 in most years). However, we also note that the proportion of data arising from rural stations was low across the entire duration although the accuracy in the last 3 years (2018–2020) in rural areas was 0.83, which could indicate improved performance with more training data from rural areas. Using a regression calibration technique, we obtained coefficients (Table S4) within land-use-based clusters and months to upscale or downscale the ensemble-averaged predictions for use in a health effects study.

### Spatiotemporal patterns of PM<sub>2.5</sub>

Clear spatiotemporal patterns exist in the annual average concentration of PM<sub>2.5</sub> across India as well as within each year (Fig. 3).

We observed higher concentrations during all years in the Indo-Gangetic plain stretching from the state of Rajasthan in the Northwest to West Bengal in the East. The southern peninsula and the mountainous regions in the North and North-east had lower concentrations. We also observed a drop in annual levels in 2020, which might be due to the reduced anthropogenic activity due to the COVID-19 lockdowns in the country.

Temporally within each year, we observed higher concentrations in the period between October and February, especially in the Indo-Gangetic plains. As temperature increases in the summer months, the levels reduce and reach minimum levels in the monsoon season (months of July and August). In the desert regions of the Northwest, higher levels were also observed in the summer months of April to June.

We examined the districtwise annual PM<sub>2.5</sub> levels and the average within-district variation in annual PM<sub>2.5</sub> within 662 districts, according to quartiles of change in population density from 2010 to 2020 (Fig. 4). Districts with larger increases in population density (above the third quartile) had higher averages as well as low variation across the years, while those below the first quartile had lower average concentrations along with higher spatial variation. In addition, we also observed an increase in spatial variation in recent years among districts with lower changes in population density. These patterns indicate that several districts with increasing population densities are also experiencing consistently higher exposure to PM<sub>2.5</sub>, which could be relevant in increasing burden from air pollution exposures for large populations.

### Scenario in select cities

Many cities in India are densely populated and often bear the burden of high levels of PM<sub>2.5</sub>. As an example of spatiotemporal variation in Indian cities, Figure 5A–E describes the annual average PM<sub>2.5</sub> levels during 2019 in the cities of Mumbai (in the West), Chandigarh (in the North), Bangalore (in the South) and Guwahati (in the East). The areas of each city were 460, 119, 708, and 4,357 km<sup>2</sup>, while the highest concentrations in these cities were 47.1, 47.5, 58.7, and 58.7 µg/m<sup>3</sup>, respectively. We observed both within and between city differences in the levels with identification of potential high-pollution areas in each city. Identification of these spatial heterogeneities makes it possible to assess short- and long-term health effects of air pollution in cohorts from these cities, using exposure averages over appropriate durations of time.

**Table 2.** A) Prediction accuracy ( $R^2$ ) of daily predictions of  $PM_{2.5}$  concentration using deep-learning, random forests, extreme gradient boosting, gradient boosting, and an ensemble-averaged model in the validation dataset, aggregated by year. B) Zone-specific prediction accuracy in the validation dataset using the predictions from the ensemble-averaged model. C) Overall, spatial, and temporal MAE, bias, and slope for the ensemble-averaged predictions of daily  $PM_{2.5}$  across 2008–2020.

A)									
Year	Deep-learning	Random forests	Gradient boosting	Extreme gradient boosting	Ensemble averaged				
					Overall	Spatial	Temporal		
2008	0.64	0.74	0.75	0.74	0.80	0.93	0.51		
2009	0.63	0.72	0.75	0.70	0.80	0.91	0.52		
2010	0.57	0.68	0.72	0.69	0.81	0.94	0.54		
2011	0.56	0.66	0.68	0.67	0.80	0.89	0.60		
2012	0.64	0.73	0.74	0.71	0.82	0.94	0.53		
2013	0.56	0.68	0.70	0.65	0.80	0.96	0.53		
2014	0.52	0.63	0.66	0.59	0.79	0.95	0.53		
2015	0.53	0.66	0.69	0.67	0.74	0.90	0.60		
2016	0.66	0.76	0.78	0.76	0.79	0.96	0.65		
2017	0.76	0.83	0.85	0.82	0.86	0.92	0.80		
2018	0.82	0.88	0.89	0.87	0.89	0.95	0.82		
2019	0.80	0.88	0.89	0.87	0.92	0.98	0.89		
2020	0.77	0.85	0.87	0.85	0.90	0.95	0.84		

B)						
Zone	Deep-learning	Random forests	Gradient boosting	Extreme gradient boosting	Ensemble averaged	
Central	0.71	0.73	0.77	0.74	0.82	
East	0.46	0.69	0.72	0.70	0.80	
North	0.74	0.84	0.86	0.84	0.87	
North-East	0.25	0.43	0.48	0.49	0.75	
South	0.26	0.42	0.43	0.39	0.63	
West	0.50	0.62	0.67	0.59	0.73	

C)						
Year	MAE			Bias	Slope	
	Overall	Spatial	Temporal			
2008	14.10	7.08	12.72	2.82	1.00	
2009	15.10	9.78	13.04	4.79	0.96	
2010	15.50	9.65	14.40	4.48	0.95	
2011	15.40	14.80	14.05	3.32	0.94	
2012	19.10	7.30	17.70	7.56	1.01	
2013	16.00	6.31	15.20	4.53	0.95	
2014	14.30	6.38	13.19	3.26	0.97	
2015	15.50	7.16	14.09	2.51	0.91	
2016	25.40	10.30	25.97	7.36	0.98	
2017	19.70	10.50	18.72	4.74	0.98	
2018	20.20	13.30	20.83	6.95	0.98	
2019	14.90	6.27	15.30	4.11	1.04	
2020	14.90	7.31	14.90	3.49	1.03	

### Scenario in an adjoining rural district

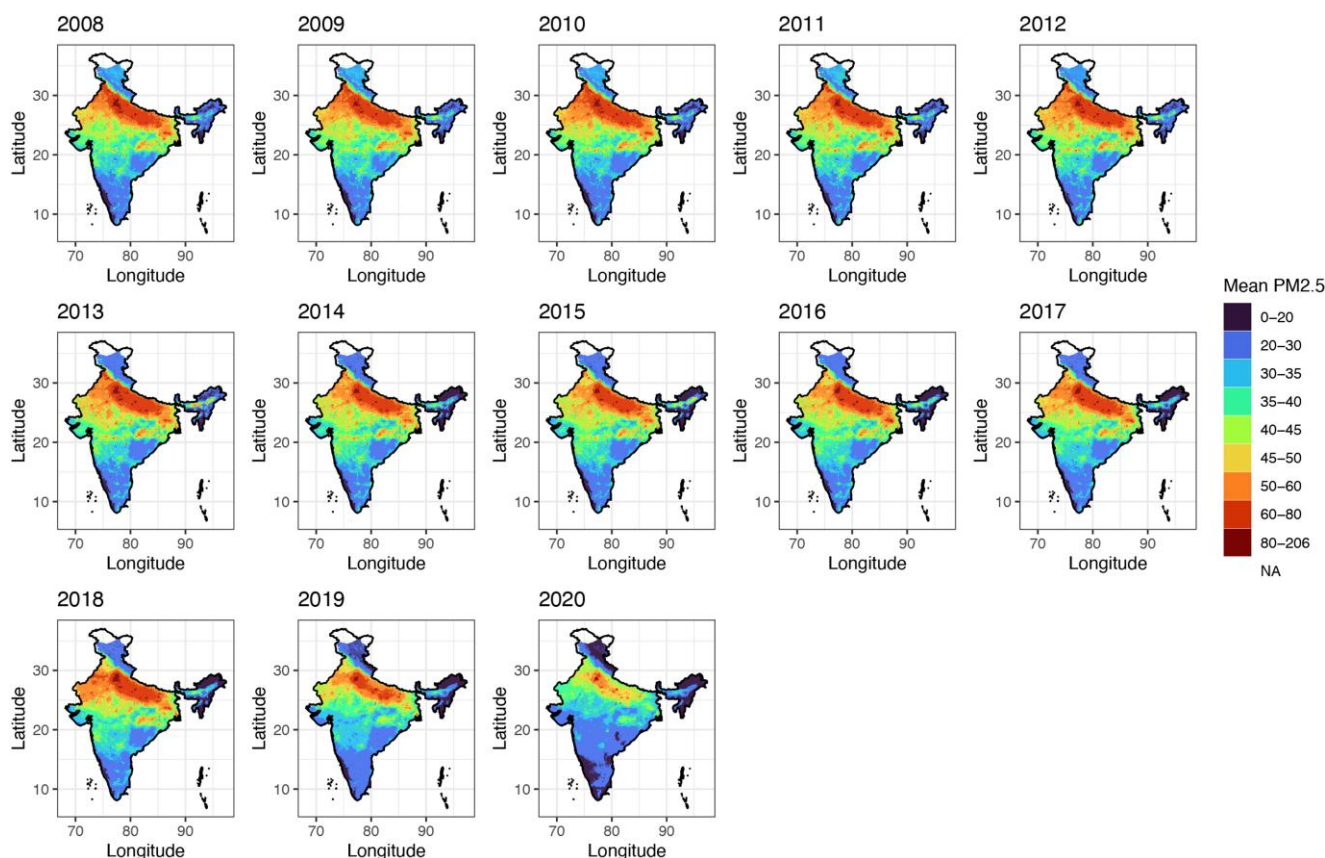
It is often assumed that periurban and rural areas are less affected by  $PM_{2.5}$ . Figure 5E shows the annual average levels during 2019 in the district of South 24 Parganas (area 6,045 km<sup>2</sup>), which neighbors the metropolitan area of Kolkata (area of 1,933 km<sup>2</sup>), a densely populated large city in the Eastern region of India. We clearly observe higher concentrations (50–80  $\mu\text{g}/\text{m}^3$ ) in close proximity to Kolkata with lower concentrations closer to the Bay of Bengal. Concentrations above 40  $\mu\text{g}/\text{m}^3$  are seen as much as 30 km south of Kolkata. This indicates how increasing population pressure and extension of the urban boundaries of major cities might influence pollution in neighboring periurban or rural areas.

### Discussion

We have developed a comprehensive prediction model to retrospectively assess daily ambient  $PM_{2.5}$  across India at a high spatial

resolution over a 13-year period, using a machine learning framework and a wide range of predictors. The findings from the model provide a useful resource to identify pollution hotspots and to study particulate matter pollution and its acute and chronic health impacts in a diverse country like India, without restricting to select geographies.

From the annual averaged predictions, we observed a slight decrease in  $PM_{2.5}$  from 2019 (highest quintile: 45.4–167  $\mu\text{g}/\text{m}^3$ ) and especially in 2020 (highest quintile: 39.9–149  $\mu\text{g}/\text{m}^3$ ), compared to 2018 (highest quintile: 52.2–186  $\mu\text{g}/\text{m}^3$ ). The effect of the COVID-19 pandemic is a possible explanation from 2020 and perhaps some of the preceding decrease may be attributed to the implementation of the National Clean Air Program (NCAP) (46) after the extreme air pollution events in 2016–2017. NCAP is a national-level strategy for a 20 to 30% reduction in  $PM_{2.5}$  and  $PM_{10}$  concentration by 2024, with 2017 as the base year for comparison. The program covers 131 nonattainment cities that did not meet the prescribed national ambient air quality standards for five



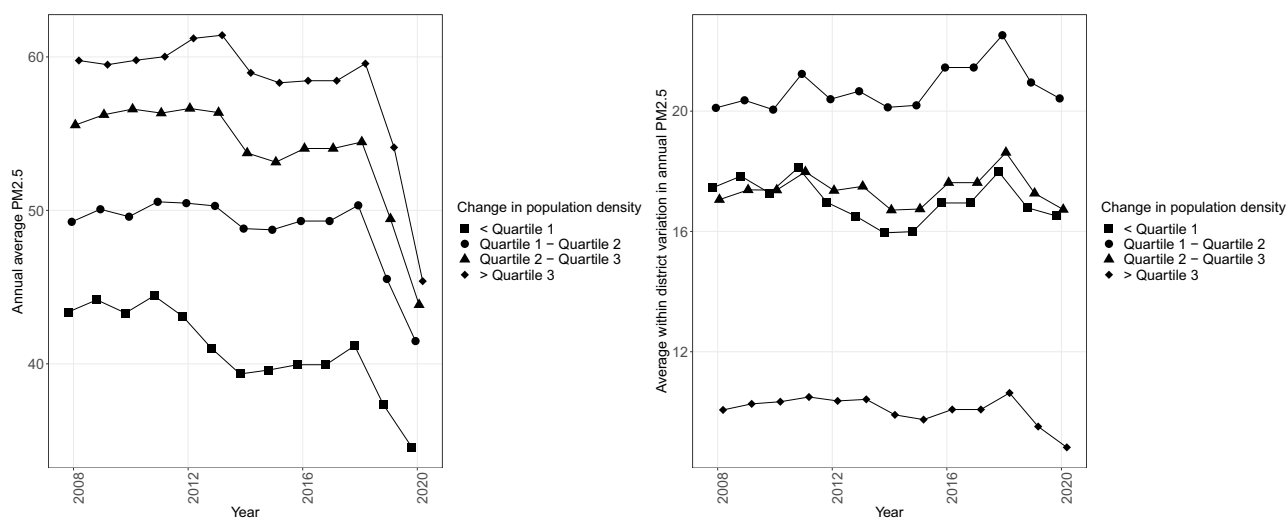
**Fig. 3.** Spatiotemporal patterns of  $PM_{2.5}$  in India: Annual average concentrations of  $PM_{2.5}$  obtained by aggregating daily ambient  $PM_{2.5}$  estimates from the ensemble-averaged model at  $1\text{ km} \times 1\text{ km}$  resolution from 2008 to 2020.

consecutive years (2011–2015). In terms of monitor density, we observed a steep increase post 2015, which could also be attributed to the NCAP program as well as the recommendation by the Central Pollution Control Board to prioritize installation of stationary monitors in urban as well as rural areas in the country.

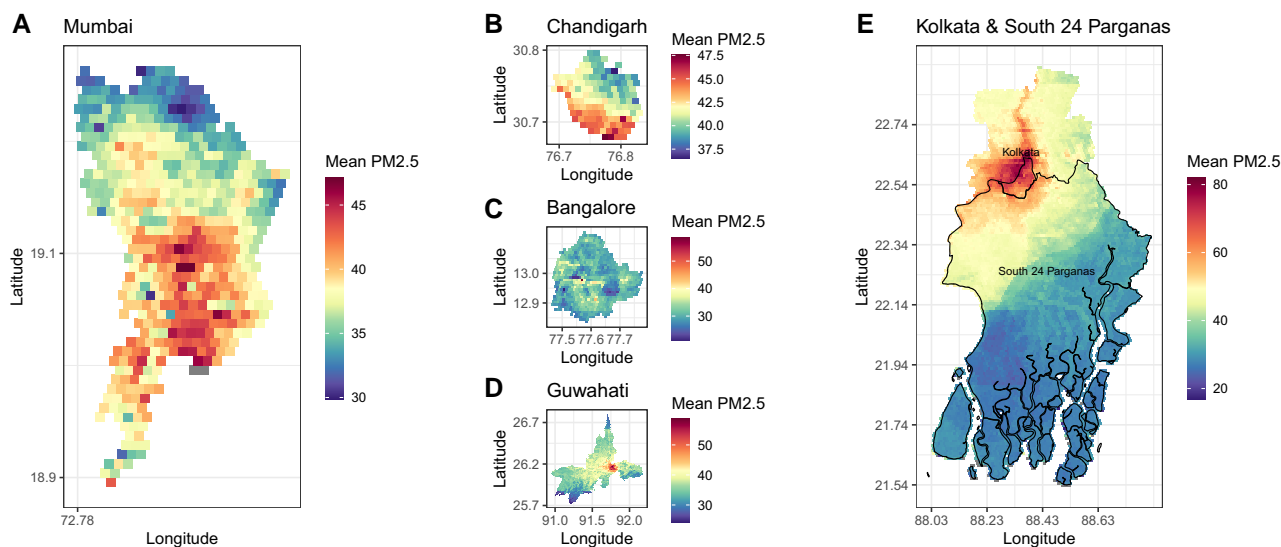
Comparing the results from our prediction model with those developed in other countries indicated that we have achieved comparable accuracy and performance with significantly less resources. The latest models for  $PM_{2.5}$  for mainland USA, which used techniques such as neural networks, random forests, and gradient boosting, combined using ensemble averaging, reported CV  $R^2$  of 0.86 for daily ambient  $PM_{2.5}$ , along with a spatial root-mean-squared-error (RMSE) of  $1.26\ \mu\text{g}/\text{m}^3$  and temporal RMSE of  $2.53\ \mu\text{g}/\text{m}^3$  (37). In Europe, ensemble modeling has been carried out in Italy (47), Great Britain (48), and Israel (49), with reported daily prediction accuracies of 0.79–0.81 (RMSE  $5.3$ – $6.6\ \mu\text{g}/\text{m}^3$ ), 0.77 (RMSE  $4.0\ \mu\text{g}/\text{m}^3$ ), and 0.87 (RMSE of  $6.1\ \mu\text{g}/\text{m}^3$ ), respectively. A recently published model from China using machine learning-based techniques reported a prediction accuracy of 0.59 with RMSE of  $29\ \mu\text{g}/\text{m}^3$  (50). However, it is important to note the difference in the number of observations (in case of the USA and China) and the smaller areas (in case of the other countries) compared to the Indian scenario.

In the case of India, there have been few attempts to develop high-resolution exposure models for ambient  $PM_{2.5}$ , each with its own limitations. In Mandal et al. (14), daily ambient  $PM_{2.5}$  was assessed at a  $1\text{ km} \times 1\text{ km}$  spatial resolution over 2010 to 2016 for the city of Delhi, which was not representative of India in terms of geography and meteorology. Further, this model used only ground monitoring data from stations around the

Delhi National Capital Region. Mhawish et al. (15) carried out a modeling exercise for the Indo-Gangetic plain using data from 2019, which was limited both temporally and spatially and also did not use spatial predictors such as land-use patterns. In a previously published satellite observation-based exposure model for the whole country, the reported prediction accuracy was 0.80 at the daily level and 0.97 at the annual level with RMSEs of  $25.7$  and  $7.2\ \mu\text{g}/\text{m}^3$ , respectively (16). Our model does noticeably better for daily values but somewhat worse for annual predictions. However, their model was trained on a 70:30 split of observations from only 120 central monitoring stations with a random cross-validation approach and no held-out data for final evaluation. Since the split was done on observations and not monitors, the spatial performance (and thereby annual and overall performance) is likely overestimated. Further, the model only relied on satellite- and reanalysis-based aerosol optical depth as a predictor of ambient  $PM_{2.5}$ , which does not account for numerous other factors such as land use and meteorology, potentially affecting the model's ability to capture local and regional variations in  $PM_{2.5}$ . Recently, an extension of this model incorporated an AOD filling algorithm and reported a fivefold cross-validation  $R^2$  of 0.92 and RMSE of  $11.8\ \mu\text{g}/\text{m}^3$  on an annual scale (51). In the modeling approach developed by Shaddick et al. (18), 82% of the global ground monitoring data used to develop the model were from 2013 and 2014 with no rural or periurban representation which makes up the majority of India. Further, the model used annual average concentrations as the modeled outcome, while using data on long-term average of  $PM_{2.5}$  from the monitoring stations in the WHO ambient air quality database. The latest iteration of this database provides only 421 annual average observations of  $PM_{2.5}$



**Fig. 4.** Mean and variation within districts by changing population density: A) Annual average  $PM_{2.5}$  by summarizing daily predictions over districts experiencing different magnitude of changes in population density from 2010 to 2020. B) Average within-district (average of districtwise standard deviations) variation categorized by quartiles of change in population density.



**Fig. 5.** Annual average  $PM_{2.5}$  in select cities: Spatial patterns of annual average  $PM_{2.5}$  (obtained by summarizing daily predictions of  $PM_{2.5}$ ) during 2019 in four urban areas, A) Mumbai (in the West), B) Chandigarh (in the North), C) Bangalore (in the South) and D) Guwahati (in the East). Additionally, the annual average for Kolkata (in the East) along with the district of South 24 Parganas is shown in (E).

from 160 unique locations with no spatial disaggregation within large cities such as Delhi.

There are several strengths of our modeling approach. A major difference between our approach with existing approaches is the use of daily average concentrations as the modeled outcome rather than annual averages. This allows us to model the short- and long-term variations which are particularly important in Indian scenarios which experience large seasonal variations as well as variations due to regional sources like agricultural crop burning. Epidemiological studies that study the acute and chronic effects of  $PM_{2.5}$  on health of individuals often require time-varying exposures of different durations starting from the day before sample collection, birth, death, or hospital admission (52, 53). Our model allows the construction of these average exposure metrics using daily concentrations rather than relying on downscaling annual averages. In addition, for our modeling purposes, we have

used daily average  $PM_{2.5}$  measurements from over 1,000 monitors across India, which we have carefully curated over the past few years. To our knowledge, this is much larger than any existing datasets for India used in other studies resulting in more robust exposure estimates. This is a major strength in comparison with the global models that rely on small sample sizes from South Asia and India, in particular (11, 17, 18).

We have used four different machine learning approaches which have specific advantages toward modeling nonlinear patterns, interaction between variables, and high dimensional interactions. We observed a poorer performance of the deep-learning algorithm compared to the gradient boosting-based approaches when data was sparse in the earlier years. However, with an increasing number of observations in the most recent years, the performances of the different methods were comparable. Further, the ensemble averaging approach using the Gaussian processes



led to an improvement in the prediction accuracy of the model in comparison to each component learner across each year. This enabled us to overcome the learner-specific deficiencies across all years to obtain accurate predictions across the entire time period, pointing to its ability to capture complex patterns across methods. Unlike previously developed models in Italy (48) or Great Britain (49), we have a better performance in terms of spatial accuracy, which is important to study the health effects of air pollution due to spatial heterogeneities. This approach can thus provide a template to model pollutants in data sparse scenarios especially in low- and middle-income countries.

As pointed out by Ravishankara et al. (11), air pollution is not exclusively an urban problem in India, with large populations in rural and periurban areas also at risk. In our dataset, we had varying degrees of data from monitors located in rural and periurban areas. The proportion of data originating from rural locations was 6% in 2008 and increased to 16% in 2020, which enhances the representativeness of our model across rural India. We specifically highlighted a scenario where high  $PM_{2.5}$  levels in densely populated urban location (Kolkata) were spilling over into periurban and rural areas surrounding it. Further, we presented an analysis of precision that highlights the need of monitoring such areas to be able to characterize exposure more precisely and facilitate the study of health effects of air pollution.

We used a cross-validation approach by leaving out monitors based on spatial clustering as well as data availability. This ensured that the model training was based on an equitable distribution across spatial clusters as well as data-heavy and data-light monitors. This was an important aspect to prevent overfitting while using multiple machine-learning algorithms, especially since most of the country's data originated from a few urban population centers in recent years. An additional feature of the model was our ability to implement regression calibration to account for potential exposure error while using these predictions in studies of health associations, which were not available in the existing exposure models. Health effect studies using exposure from prediction models lack gold standard exposure measurements, potentially introducing bias in the associations between  $PM_{2.5}$  and health. Hence the availability of an exposure error-corrected prediction, leveraging a validation dataset, is an important resource for future health studies.

Despite the efforts to develop a comprehensive model, there are limitations to our model. First, we used variables that were available for all grids across the country. However, there may be sources of  $PM_{2.5}$  that are specific to certain neighborhoods, for example, shopping malls and commercial areas in urban areas, for which information was not available nationally. Also, we did not have access to any validated source for spatiotemporally varying data on traffic flow, which is a major contributor of  $PM_{2.5}$ . Inclusion of these predictors to build hyperlocal models of  $PM_{2.5}$  may increase the accuracy of our model. Second, we observe a poorer performance of the model in the early years, most likely due to the sparseness of ground monitoring data. However, we used a  $PM_{2.5}$  and  $PM_{10}$  calibration model to fill in gaps in the  $PM_{2.5}$  ground monitoring data. Third, we observed large gaps in the MAIAC AOD data over the country and used a deep-learning-based imputation to fill the gaps. However, the coarse resolution and inherent uncertainty in the CAMS AOD as well as meteorological variables may introduce biases in the imputed AOD. Further, our model is not equipped to fully capture extreme events due to localized sources such as biomass burning during winter in Northern India, which often leads to spikes in air pollution. While we did use MODIS fire and global fire

emissions database (GFED) carbon emissions from fires in our models, these variables do not fully capture short-term variations in biomass burning. However, in studies of long-term health effects of air pollution in a population setting, these limitations should not hamper our inferences. Further, access to above-mentioned predictors can help facilitate updating the model for the future years. The existing model will be updated for following years using additional ground monitoring data and for select cities, models will be updated to obtain predictions at a finer spatial resolution of  $200\text{ m} \times 200\text{ m}$ .

## Conclusion

In this article, we presented a comprehensive state-of-the-science resource to assess daily average  $PM_{2.5}$  concentration at fine spatio-temporal resolution across a large, diverse, and populous country over a time period of 13 years. This unique model is an important resource to fill the gaps in air pollution epidemiology research in India. The modeled  $PM_{2.5}$  can be effectively leveraged to study associations with a range of health outcomes across urban, periurban, and rural India.

## Acknowledgments

The authors acknowledge Dr Cathryn Tonne (IS-Global, Barcelona) and Prof. Tirthankar Banerjee (Banaras Hindu University, Varanasi, India) for providing additional ground monitoring data for  $PM_{2.5}$ .

## Supplementary Material

Supplementary material is available at PNAS Nexus online.

## Funding

This research was conducted with funding from the Swedish Research Council for Sustainable Development ("FORMAS") under the project titled "Consortium for Climate, Health, and Air Pollution Research in India (CHAIR-India)" (Award number: 2020-00446).

## Author Contributions

S.M., I.K., D.P., P.P., P.L., and J.S. formulated the paper and modeling approaches. S.M., A.R., J.S.M., K.J.L., G.K.W., S.D., A.N., A.D., P.S., and S.J. downloaded and processed data. H.A. and A.N. provided certain datasets on predictors and ground monitoring. A.R. and H.A. prepared the figures for the paper. C.V. provided the emissions inventories for India. S.M., A.R., and J.M. carried out the modeling exercise with inputs from I.K., H.A., and J.S. S.M. formulated the draft with assistance from all coauthors. P.L. was awarded the grant that provided the funding for this exercise.

## Data Availability

Data on  $PM_{2.5}$  predictions and codes to develop the machine learning models will be shared in a public repository as detailed in the project proposal. Curated and processed data on predictors will be shared on request since the project is ongoing and more publications are planned as part of the project using these data. As part of the publications policy in the grant, an academic agreement would be required to share the data. Contact Siddhartha Mandal, [siddhartha@ccdcindia.org](mailto:siddhartha@ccdcindia.org) to request data access.

## References

- 1 Tapia VL, et al. 2020. Association between maternal exposure to particulate matter (PM<sub>2.5</sub>) and adverse pregnancy outcomes in Lima, Peru. *J Expo Sci Environ Epidemiol*. 30(4):689–697.
- 2 Fang J, et al. 2020. Prenatal PM<sub>2.5</sub> exposure and the risk of adverse births outcomes: results from project ELEFANT. *Environ Res*. 191:110232.
- 3 Kim H, et al. 2017. Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900845 person-years of follow-up. *J Am Heart Assoc*. 6(11):e007170.
- 4 He D, et al. 2017. Association between particulate matter 2.5 and diabetes mellitus: a meta-analysis of cohort studies. *J Diabetes Invest*. 8(5):687–696.
- 5 Brunekreef B, et al. 2009. Effects of long-term exposure to traffic-related air pollution on respiratory and cardiovascular mortality in The Netherlands: the NLCS-AIR study. *Res Rep Health Eff Inst*. 139:5–71; discussion 73–89.
- 6 Grande G, et al. 2020. Association between cardiovascular disease and long-term exposure to air pollution with the risk of dementia. *JAMA Neurol*. 77(7):801–809.
- 7 Rajagopalan S, Al-Kindi SG, Brook RD. 2018. Air pollution and cardiovascular disease: JACC state-of-the-art review. *J Am Coll Cardiol*. 72(17):2054–2070.
- 8 Johnson NM, et al. 2021. Air pollution and children's health—a review of adverse effects associated with prenatal exposure from fine to ultrafine particulate matter. *Environ Health Prev Med*. 26:1–29.
- 9 Pandey A, et al. 2021. Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. *Lancet Planet Health*. 5(1):e25–e38.
- 10 Balakrishnan K, et al. 2019. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the global burden of disease study 2017. *Lancet Planet Health*. 3(1):e26–e39.
- 11 Ravishankara AR, et al. 2020. Outdoor air pollution in India is not only an urban problem. *Proc Natl Acad Sci U S A*. 117(46):28640–28644.
- 12 Brauer M, et al. 2019. Examination of monitoring approaches for ambient air pollution: a case study for India. *Atmos Environ*. 216:116940.
- 13 Hoek G. 2017. Methods for assessing long-term exposures to outdoor air pollutants. *Curr Environ Health Rep*. 4:450–462.
- 14 Mandal S, et al. 2020. Ensemble averaging based assessment of spatiotemporal variations in ambient PM<sub>2.5</sub> concentrations over Delhi, India, during 2010–2016. *Atmos Environ*. 224:117309.
- 15 Mhawish A, et al. 2020. Estimation of high-resolution PM<sub>2.5</sub> over the Indo-Gangetic plain by fusion of satellite data, meteorology, and land use variables. *Environ Sci Technol*. 54(13):7891–7900.
- 16 Dey S, et al. 2020. A satellite-based high-resolution (1-km) ambient PM<sub>2.5</sub> database for India over two decades (2000–2019): applications for air quality management. *Remote Sens (Basel)*. 12(23):3872.
- 17 Hammer MS, et al. 2020. Global estimates and long-term trends of fine particulate matter concentrations (1998–2018). *Environ Sci Technol*. 54(13):7879–7890.
- 18 Shaddick G, et al. 2018. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *J R Stat Soc Ser C Appl Stat*. 67(1):231–253.
- 19 Kloog I, et al. 2013. Long- and short-term exposure to PM<sub>2.5</sub> and mortality: using novel exposure models. *Epidemiology*. 24(4):555–561.
- 20 Ma Z, et al. 2022. Short-term effects of different PM<sub>2.5</sub> ranges on daily all-cause mortality in Jinan, China. *Sci Rep*. 12(1):5665.
- 21 Hurtado-Díaz M, et al. 2021. Prenatal PM<sub>2.5</sub> exposure and neurodevelopment at 2 years of age in a birth cohort from Mexico City. *Int J Hyg Environ Health*. 233:113695.
- 22 Prabhakaran D, et al. 2020. Exposure to particulate matter is associated with elevated blood pressure and incident hypertension in urban India. *Hypertension*. 76(4):1289–1298.
- 23 Mandal S, et al. 2023. PM<sub>2.5</sub> exposure, glycemic markers and incidence of type 2 diabetes in two large Indian cities. *BMJ Open Diabetes Res Care*. 11(5):e003333.
- 24 Spiegelman D. 2013. Regression calibration in air pollution epidemiology with exposure estimated by spatio-temporal modeling. *Environmetrics*. 24(8):521–524.
- 25 Chen T, et al. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, p. 1–4.
- 26 Lyapustin A, et al. 2018. MODIS collection 6 MAIAC algorithm. *Atmos Meas Tech*. 11(10):5741–5765.
- 27 Giglio L, Schroeder W, Justice CO. 2016. The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sens Environ*. 178:31–41.
- 28 Elvidge CD, et al. 2017. VIIRS night-time lights. *Int J Remote Sens*. 38(21):5860–5879.
- 29 Copernicus Sentinel-5P (processed by ESA). 2018. TROPOMI Level 2 Nitrogen Dioxide total column products. Version 01. European Space Agency. doi:10.5270/S5P-s4ljg54.
- 30 Hersbach H, et al. 2020. The ERA5 global reanalysis. *Q J R Meteorol Soc*. 146(730):1999–2049.
- 31 Doxsey-Whitfield E, et al. 2015. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Pap Appl Geogr*. 1(3):226–234.
- 32 NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team. 2019. ASTER Global Digital Elevation Model V003 [Data set]. NASA EOSDIS Land Processes DAAC. doi:10.5067/ASTER/ASTGTM.003.
- 33 Venkataraman C, et al. 2018. Source influence on emission pathways and ambient PM<sub>2.5</sub> pollution over India (2015–2050). *Atmos Chem Phys*. 18(11):8017–8039.
- 34 Randerson JT, et al. 2015. Global fire emissions database, Version 4 (GFEDv4). Oak Ridge (TN): ORNL DAAC. doi:10.3334/ORNLDAAC/1293.
- 35 Peuch V-H, et al. 2022. The copernicus atmosphere monitoring service: from research to operations. *Bull Am Meteorol Soc*. 103(12):E2650–E2668.
- 36 Kloog I, et al. 2015. Estimating daily PM<sub>2.5</sub> and PM<sub>10</sub> across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmos Environ*. 122:409–416.
- 37 Di Q, et al. 2019. An ensemble-based model of PM<sub>2.5</sub> concentration across the contiguous United States with high spatio-temporal resolution. *Environ Int*. 130:104909.
- 38 Bengio Y. 2016. *Deep learning*. Cambridge (USA): MIT Press.
- 39 Atkinson PM, Tatnall AR. 1997. Introduction neural networks in remote sensing. *Int J Remote Sens*. 18(4):699–709.
- 40 Breiman L. 2001. Random forests. *Mach Learn*. 45:5–32.
- 41 Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 29(5):1189–1232.
- 42 Vecchia AV. 1988. Estimation and model identification for continuous spatial processes. *J R Stat Soc Ser B (Methodol)*. 50(2):297–312.
- 43 Guinness J. 2018. Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*. 60(4):415–429.
- 44 Feng Y, et al. 2023. Measurement error correction for ambient PM<sub>2.5</sub> exposure using stratified regression calibration: effects on all-cause mortality. *Environ Res*. 216:114792.
- 45 Aschenbruck R, Szepannek G, Wilhelm AF. 2022. Imputation strategies for clustering mixed-type data with missing values. *J Classif*. 40(1): 1–23.

- 
- 46 Ganguly T, Selvaraj KL, Guttikunda SK. 2020. National Clean Air Programme (NCAP) for Indian cities: review and outlook of clean air action plans. *Atmos Environ.* 8:100096.
  - 47 Shtein A, et al. 2019. Estimating daily PM<sub>2.5</sub> and PM<sub>10</sub> over Italy using an ensemble model. *Environ Sci Technol.* 54(1):120–128.
  - 48 Schneider R, et al. 2020. A satellite-based spatio-temporal machine learning model to reconstruct daily PM<sub>2.5</sub> concentrations across Great Britain. *Remote Sens (Basel).* 12(22):3803.
  - 49 Shtein A, et al. 2018. Estimating daily and intra-daily PM<sub>10</sub> and PM<sub>2.5</sub> in Israel using a spatio-temporal hybrid modeling approach. *Atmos Environ.* 191:142–152.
  - 50 Xiao Q, et al. 2018. An ensemble machine-learning model to predict historical PM<sub>2.5</sub> concentrations in China from satellite data. *Environ Sci Technol.* 52(22):13260–13269.
  - 51 Katoch V, et al. 2023. Addressing biases in ambient PM<sub>2.5</sub> exposure and associated health burden estimates by filling satellite AOD retrieval gaps over India. *Environ Sci Technol.* 57(48):19190–19201.
  - 52 Lee S, et al. 2019. Short-term PM<sub>2.5</sub> exposure and emergency hospital admissions for mental disease. *Environ Res.* 171:313–320.
  - 53 Yitshak-Sade M, et al. 2021. PM<sub>2.5</sub> and hospital admissions among medicare enrollees with chronic debilitating brain disorders. *Sci Total Environ.* 755:142524.