https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

Open Access Earth System Science Data Discussions

Reconstructing long-term (1980-2022) daily ground particulate matter datasets in India (LongPMInd)

Shuai Wang[1], Mengyuan Zhang[1], Hui Zhao[2], Peng Wang[3,4], Sri Harsha Kota[5], Qingyan Fu[6], Cong Liu[7], Hongliang Zhang[1,4,8*]

[1]Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, China
[2]School of Resources and Environmental Engineering, Jiangsu University of Technology, Changzhou 213001, China
[3]Department of Atmospheric and Oceanic Sciences, and Institute of Atmospheric Sciences, Fudan University, Shanghai, 200438, China
[4]IRDR ICoE on Risk Interconnectivity and Governance on Weather/Climate Extremes Impact and Public Health, Fudan University, Shanghai, China
[5]Department of Civil Engineering, Indian Institute of Technology, Delhi, 110016, India
[6] Shanghai Academy of Environmental Sciences, Shanghai 200003, China
[7]School of Public Health, Fudan University, Shanghai, 200032, China
[8]Institute of Eco-Chongming (IEC), Shanghai 200062, China

Correspondence to: Hongliang Zhang (zhanghl@fudan.edu.cn)

Abstract. Severe airborne particulate matter (PM, including $PM_{2.5}$ and PM_{10}) pollution in India has caused widespread concern. Accurate PM datasets are fundamental for scientific policymaking and health impact assessment, while surface observations in India are limited due to scarce sites and uneven distribution. In this work, a simple structured, efficient, and robust model based on the Light Gradient Boosting Machine (LightGBM) was developed to fuse multi-source data and estimate long-term (1980-2022) historical daily ground PM datasets in India (LongPMInd). The LightGBM model shows good accuracy with out-of-sample, out-of-site, and out-of-year cross-validation CV test R^2 of 0.77, 0.70, and 0.66, respectively. Small performance gaps between $PM_{2.5}$ training and testing (delta RMSE of 1.06, 3.83, and 7.74 μg m^{-3}) indicate low overfitting risks. With great generalization ability, the open-accessible, long-term, and high-quality daily $PM_{2.5}$ and PM_{10} products were then reconstructed (10 km, 1980-2022). It shows that India has experienced severe PM pollution in the Indo-Gangetic Plain (IGP), especially in winter. PM concentrations significantly increased ($p<0.05$) in most regions since 2000 (0.34 μg m^{-3} $year^{-1}$). The turning point occurred in 2018 when the Indian government launched the National Clean Air Program, $PM_{2.5}$ concentrations declined in most regions (- 0.78 μg m^{-3} $year^{-1}$) during 2018-2022. Severe $PM_{2.5}$ pollution caused continuous increased attributable premature mortalities, from 0.73 (95 % CI: 0.65-0.80) million in 2000 to 1.22 (95 % CI: 1.03-1.41) million in 2019, particularly in the IGP, where attributable mortality increased from 0.36 to 0.60 million. The LongPMInd datasets have the potential to support multi-applications of air quality management, public health, and climate change. The daily and monthly $PM_{2.5}$ and PM_{10} datasets are publicly accessible at https://doi.org/10.5281/zenodo.10073944 (Wang et al., 2023a).

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

1 Introduction

Airborne particulate matter (PM, including $PM_{2.5}$ with diameters < 2.5 μm and PM_{10} with diameters < 10 μm) not only impacts climate by changing radiation budgets but also has significant adverse effects on human health(Murray et al., 2020; Wang et al., 2012; Yang et al., 2016). India is one of the most populous countries, with severe PM pollution resulting from rapid economic development and industrialization over the last few decades. Exposure to $PM_{2.5}$ has become one of the leading causes of health burden in India, including heart disease, stroke, lung cancer, and premature death(Dandona et al., 2017; Pandey et al., 2021).

Accurate datasets of ground PM concentration are prerequisites for evidence-based policymaking and health impact assessments. The Central Pollution Control Board (CPCB) of India has established and maintained ground-based monitoring networks with ~335 continuous ambient air quality monitoring stations (CAAQMS) currently. However, these monitoring sites are unevenly distributed (mainly located in urban, residential, and industrial areas), with limited number of sites (monitoring density: ~0.6 sites per million population) (Brauer et al., 2019), and many cities even have no monitoring sites (Dey et al., 2020; Martin et al., 2019). Therefore, the surface observations alone are not sufficient to support air quality management, especially on a regional scale (Pant et al., 2019).

Several studies have explored different methods to estimate ground $PM_{2.5}$ concentrations in India. Bali et al. (Bali et al., 2021) estimated total $PM_{2.5}$ through empirical coefficients using MERRA2, while these coefficients vary with geographic location and pollution scenarios, which makes the estimation potentially unreliable. Chowdhury et al. (Chowdhury et al., 2019) used the $PM_{2.5}$ – AOD (aerosol optical depth) equation method to estimate $PM_{2.5}$ concentrations in Delhi, however, AOD satellite data suffers from significant non-random misses, especially during cloud cover and hazy polluted days, so it is difficult to derive a spatiotemporal full-coverage PM dataset (Bai et al., 2022; Wang et al., 2023d).

Multi-source data fusion approaches coupled with artificial intelligence technology have been increasingly used to extend the record of air pollutants like $PM_{2.5}$, including satellite observations, meteorological fields, and emission inventories (Wang et al., 2023d; Wei et al., 2021; Ren et al., 2022b; Katoch et al., 2023). Tree-based machine learning (ML) models typically outperform deep learning approaches in tabular data (e.g., air pollutant observation datasets), and thus have been widely developed (Grinsztajn et al., 2022). Wei et al. (Wei et al., 2021) reconstructed long-term high-quality $PM_{2.5}$ data records in China by fusing satellite, meteorological, and emission data using a spatiotemporal extra tree (STET) model. Xue et al. (Xue et al., 2020) estimated ground ozone concentration in China by ML-based data-fusion methods. Sayeed et al. (Sayeed et al., 2022) improved the $PM_{2.5}$ concentration in the continental United States using Random Forest (RF) approach coped with meteorology and aerosol species of MERRA-2. Some studies have demonstrated the feasibility of tree-based model to estimate $PM_{2.5}$ concentrations in India (Kumar et al., 2023; Dhandapani et al., 2023; Bali et al., 2019). However, it is challenging to establish long-term, full-coverage, high accuracy, open-source PM data products in India due to insufficient model robustness and implementation capacity (Dey et al., 2020; Kumar et al., 2023).

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

To improve performance, previous models usually have high complexity, such as numerous trees and leaf nodes (Zhang et al., 2021; Huang et al., 2021). This practice raises the requirement of computational resources and is prone to overfitting, leading to a large gap between the performance of the training and testing(Zhang et al., 2021; Jabbar and Khan, 2015; Ying, 2019). Wang et al.(Wang et al., 2023b) compared a simple linear model with the tree-based XGboost model and found that XGboost was much slower (> 1000 %) and suffered a higher overfitting risk. Therefore, it is necessary to minimize model complexity to avoid overfitting.

In this work, a simple structured, efficient, and robust model based on the Light Gradient Boosting Machine (LightGBM) was developed to estimate PM concentration. Three cross-validation methods and separate test datasets were designed to evaluate model performance. Long-term (1980-2022) and open-source datasets with a spatial resolution of 10 km of $PM_{2.5}$ and PM_{10} in India were then generated, and the mortalities due to $PM_{2.5}$-induced diseases were also estimated. The datasets could help with pollution formation analysis, assessment of PM health risks, and air quality management in India.

2 Materials and methods

2.1 Data sources

Table 1 shows the multisource datasets used in this study. Ground observations of $PM_{2.5}$ and PM_{10} during 2018-2022 in India were collected from the CPCB air quality monitoring network (www.cpcb.nic.in). The location of monitoring sites is shown in Fig. S1. Observations data less than 0.01 % and larger than 99.99 % were excluded. The fifth generation ECMWF atmospheric reanalysis datasets ERA5-Land in 1980-2022 were collected, and several meteorological factors with high relative importance are included (Table 1). Datasets of Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) in 1980-2022 were also collected, including aerosol optical depth and aerosol components and precursors (black carbon, organic carbon, sulfate, dust, and SO_2).

Table 1: Summary of the ERA5, MERRA2, and ground observation data used in this study.

Type	Variable	Description	Spatial Resolution	Temporal Resolution
ERA5	SSRD	Surface solar radiation	0.1° × 0.1°	Hourly
	BLH	Boundary layer height	0.25°× 0.25°	Hourly
	EVAP	Evaporation	0.1°× 0.1°	Hourly
	TEMP2	2m air temperature	0.1°× 0.1°	Hourly
	DEWP2	2m dewpoint temperature	0.1°× 0.1°	Hourly
	SP	Surface pressure	0.1°× 0.1°	Hourly
	TPREC	Total precipitation	0.1°× 0.1°	Hourly
	TCLOUD	Total cloud cover	0.25°× 0.25°	Hourly
	UWIND10	10m u component of wind	0.1°× 0.1°	Hourly

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

	VWIND10	10m v component of wind	0.1°× 0.1°	Hourly
MERRA2	BCSMASS	Black carbon surface mass concentration	0.5 °× 0.625 °	Hourly
	OCSMASS	Organic carbon surface mass concentration	0.5 °× 0.625 °	Hourly
	DUSMASS25	Dust– $PM_{2.5}$ surface mass concentration	0.5 °× 0.625 °	Hourly
	DUSMASS	Dust surface mass concentration	0.5 °× 0.625 °	Hourly
	SO2SSMASS	Sulfur dioxide surface mass concentration	0.5 °× 0.625 °	Hourly
	SO4SMASS	Sulfate surface mass concentration	0.5 °× 0.625 °	Hourly
	TOTEXTTAU	Total aerosol extinction [550 nm]	0.5 °× 0.625 °	Hourly
Observation	$PM_{2.5}$, PM_{10}	Particulate matter	Point	Hourly

2.2 Model building

In this study, LightGBM (Ke et al., 2017), an efficient Gradient Boosting Decision Tree (GBDT), was used to estimate $PM_{2.5}$ and PM_{10}, which has been proven to be accurate, fast, and robust in our previous studies (Wang et al., 2023b; Wang et al., 2023c). Grid search cross-validation (CV) method were used to select the optimal hyperparameters. An algorithm for hyperparameter selection (SI: Algorithm 1) was designed to ensure the model's generalization ability. Loop to increase the model complexity (e.g., number of trees), ending the loop and returning the hyperparameters when the model predicted RMSE does not decrease significantly (< 0.01) or the difference between training and predicted RMSE does not increase significantly (< 0.05). Features were selected based on the relative importance. Ten meteorological features, six emission-related features, and total aerosol extinction were used to train the LightGBM and estimate PM concentrations (Fig. 1). The meteorological and emission features contributed 64% and 31% to the $PM_{2.5}$ prediction.

Three independent CV methods and three metrics (coefficient of determination: R^2, root mean square error: RMSE, and mean absolute error: MAE) were designed to evaluate the model's spatiotemporal predictive power. The first is out-of-sample CV, where the dataset is randomly divided into 10 subsets, one of which is taken in turn for testing, and the remaining 9 subsets are used for training, which is repeated 10 times and averaged. The second is out-of-site CV, which is similar to the out-of-sample CV, but the dataset is randomly divided by site. This method can measure the model's spatial predictive power. The third method is interannual out-of-year CV, which sequentially takes one year of data for testing and the rest for training. This approach can measure the model’s predictive power for the years with no observations. Besides, observations in January-June 2023 were used as a separate test set, and these data were not involved in any of the training and hyperparameter selection processes.

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

Open Access Earth System Science Data Discussions

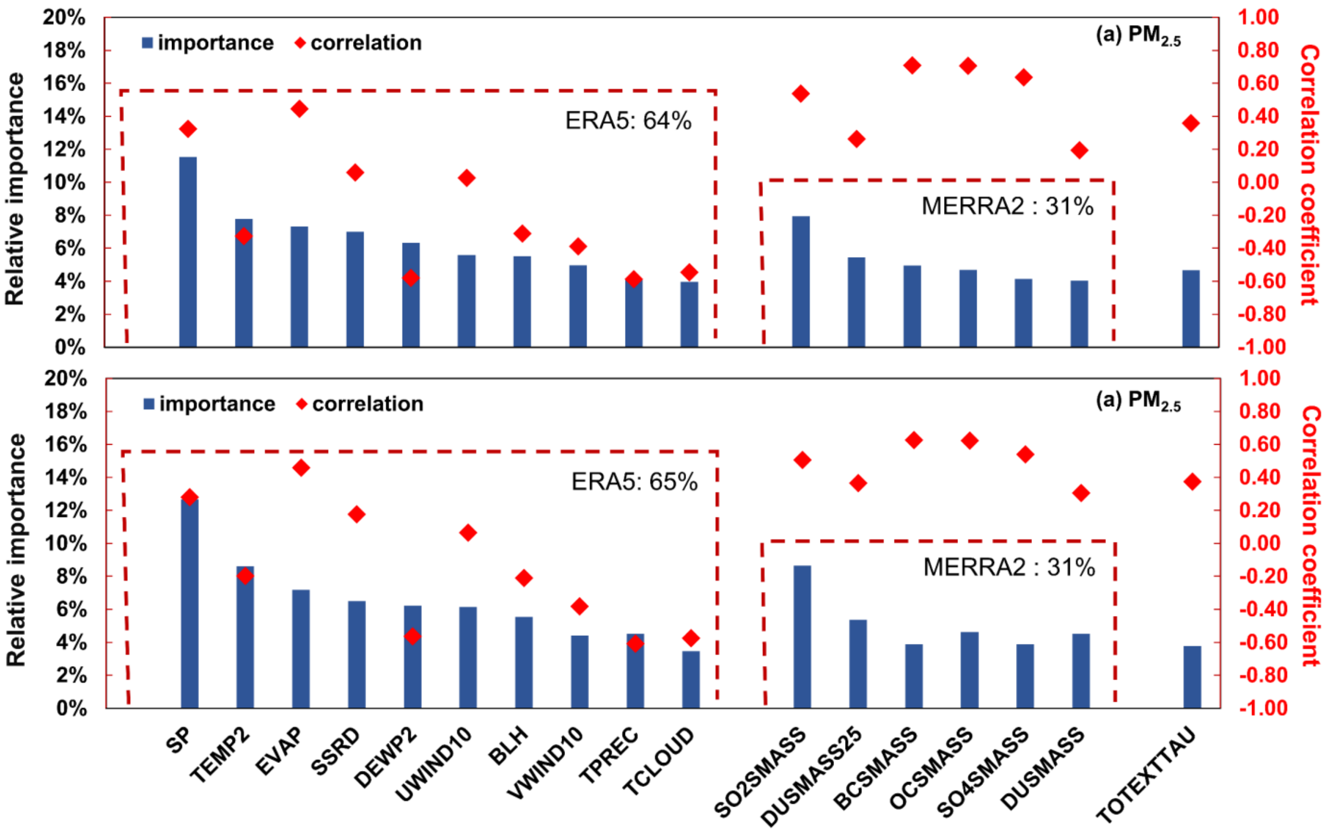


Figure 1: Relative importance and correlation coefficient for the $PM_{2.5}$ and PM_{10} estimates models. Description of the features is shown in Table 1.

2.3 Mortality estimation.

According to the GBD 2019 study(Murray et al., 2020; Vos et al., 2020), annual average concentrations were used to assess long-term exposure to $PM_{2.5}$, and premature deaths were assessed using the following equation:

$$M_{y,i,j} = \frac{RR_j(C_{y,i}) - 1}{RR_j(C_{y,i})} \times P_{y,i} \times I_{y,j}$$

Where, $M_{y,\,i}$ represents the mortality attributable to cause j due to long-term $PM_{2.5}$ exposure in year y in region i. $RR(C_{y,\,i})$ represents the relative risk of cause j for year y in region i. $P_{y,\,i}$ represents the population j in year y in region i, and I_y represents the baseline mortality in year y.

$PM_{2.5}$ exposure-related deaths due to ischemic heart disease (CVD_IHD), chronic stroke (CVD_stroke), obstructive pulmonary disease (RESP_COPD), lung cancer (NEO_LUNG), lower respiratory infections (LRI), and diabetes mellitus type II (T2_DM) were estimated. The gridded population data was obtained from the WorldPop datasets (https://www.worldpop.org). Annual baseline mortality (2000-2019) and risk of cause-specific deaths at different $PM_{2.5}$ levels exposure was obtained from GBD 2019. The health terminals health effects of $PM_{2.5}$ are in the range of 2.4 to 5.9 μg m^{-3}.

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

3 Results

3.1 Long-term India $PM_{2.5}$ dataset

Applying the trained LightGBM model to the large input dataset constructed for the years 1980 - 2022, the long-term high-quality daily $PM_{2.5}$ and PM_{10} products of India (LongPMInd) are reconstructed. Table 2 summarizes the basic information about the dataset, the data is provided in NetCDF format with a spatial resolution of 10 km. LongPMInd dataset to the best of our knowledge is the first open-source, longest term (i.e. 1980-2022) and relatively high accuracy dataset covering the entire India. The daily, monthly, and yearly $PM_{2.5}$ and PM_{10} datasets are publicly available at https://doi.org/10.5281/zenodo.10073944 (Wang et al., 2023a).

Table 2: Summary of the LongPMInd dataset

Data description	LongPMInd dataset
Data type	Gridded
File format	NetCDF
Specie	$PM_{2.5}$, PM_{10}
Spatial reference	WGS 84
Horizontal resolution	0.1° × 0.1° (≈ 10 km × 10 km)
Horizontal coverage	India, [60° E, 100° E], [5.0° N, 40.0° N]
Temporal resolution	Daily, monthly, and yearly
Temporal coverage	1980-2022

3.2 Model performance

Table 3 shows the training and testing results of out-of-sample CV, out-of-site CV, and out-of-year CV for daily $PM_{2.5}$ and PM_{10}. Overall, the model shows good accuracy with out-of-sample CV R^2 of 0.77, 0.76, and RMSE of 29.57, 51.63 μg m^{-3} for daily $PM_{2.5}$ and PM_{10}. Monthly predictions show better performance with out-of-sample CV R^2 of 0.87, 0.86, and RMSE of 17.65, 31.26 μg m^{-3} for monthly $PM_{2.5}$ and PM_{10}. More importantly, out-of-sample CV results of training and testing showed small accuracy gaps with RMSE and MAE of 1.06 (4 %) and 0.51 (3 %) μg m^{-3} for $PM_{2.5}$, and 1.52 (3 %) and 0.9 (3 %) μg m^{-3} for PM_{10} reflecting good generalization ability. Out-of-site CV measures the model's predictive ability for unobserved areas. The spatially validated R^2 and RMSE for $PM_{2.5}$ and PM_{10} were 0.70, 0.65, and 31.73, 51.37 μg m^{-3}, respectively, indicating model's ability to fill the unobserved areas accurately. The small performance gap between out-of-site CV training and testing also reflects good spatial generalization ability. Observations before 2018 are limited due to the number and quality of sites. Out-of-year CV was used to evaluate LightGBM prediction performance, which was conducted by sequentially taking one-year data for testing and the rest for training. The model accuracy predicts historical $PM_{2.5}$ and PM_{10} concentrations, with small RMSE (35.35 and 60.65 μg m^{-3}) and MAE(21.54 and 40.74 μg m^{-3}), suggesting that the models are reliable to reconstruct the long-term historical dataset of $PM_{2.5}$ and PM_{10} in India. Notably, most predictions are consistent with observations, with most

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

Open Access Earth System Science Data Discussions

data samples evenly distributed around the 1：1 line (Fig. S2), but with the underestimation for high PM levels and overestimation for low PM levels (slopes:0.75 and 0.74, intercepts: 16.45 and 35.79 μg m^{-3} for daily $PM_{2.5}$ and PM_{10} predictions). Monthly predictions show better agreement with observations with slopes of 0.84 and 0.82, and intercepts of 10.26 and 23.53 μg m^{-3} for monthly $PM_{2.5}$ and PM_{10}. The under- and over-estimation indicate potential unreliability of model predictions for extreme pollution and extreme clean days. This can be attributed to the small proportion of data records for extreme pollution and clean days.

Observations from January to June in 2023 were used for testing, which were not involved in any training or hyperparameter selecting processes (Fig. S3 and Table S1). Six representative regions were selected for the analysis including Delhi and Uttar Pradesh (IGP region), Gujarat (Western India region), Madhya Pradesh (Central India region), West Bengal (Eastern India region), and Andhra Pradesh (Southern India region). The model shows accurate prediction ability with RMSE of 33.58 and 64.25 μg m^{-3} for $PM_{2.5}$ and PM_{10} respectively in India. The model can capture the decreasing trend of PM concentration from January to June in different regions of India but with some biases, e.g., overestimation of $PM_{2.5}$ in Uttar Pradesh on 8 January; and underestimation of haze pollution in Gujarat on 19 February. The large RMSE of $PM_{2.5}$ prediction in Uttar Pradesh (32.72 μg m^{-3}) could be attributed to the complexity of pollution causes in the region as well as insufficient observation data. The small RMSE (8.34 μg m^{-3}) of $PM_{2.5}$ prediction in Andhra Pradesh can be related to the light haze pollution and small fluctuation of $PM_{2.5}$ concentration.

Table 3: Training and testing results of out-of-sample CV, out-of-site CV, and out-of-year CV for daily $PM_{2.5}$ and PM_{10} (2018-2022). RSME and MAE unit: μg m^{-3}.

Spec	Type	R^2		RMSE (μg m^{-3})		MAE (μg m^{-3})	
		Test	Train	Test	Train	Test	Train
$PM_{2.5}$	out-of-sample	0.77	0.79	29.57	28.51	18.76	18.25
	out-of-site	0.70	0.79	31.73	27.90	20.32	17.78
	out-of-year	0.66	0.79	35.35	27.61	21.54	17.61
PM_{10}	out-of-sample	0.76	0.77	51.63	50.11	35.42	34.52
	out-of-site	0.65	0.77	57.37	49.42	39.92	33.94
	out-of-year	0.66	0.78	60.65	49.06	40.74	33.72

3.3 Spatial and temporal trends

First, spatial patterns of $PM_{2.5}$ and PM_{10} are analyzed (Fig. S4 and S5). The Indo-Gangetic Plain (IGP) and western arid region show high levels of $PM_{2.5}$ and PM_{10}, especially for years after 2000. Low PM concentrations were observed in south India. The high terrain in the north and south IGP is unfavorable for pollutant dispersion. Intense human activities in IGP (population > 700 million) emit large amounts of primary PM and gas pollutants (SO_2 and nitrogen oxide) coupled with unfavorable dispersion conditions leading to severe PM pollution (Dey et al., 2020; Maheshwarkar et al., 2022). Both $PM_{2.5}$ and PM_{10}

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

concentrations show north-to-south (high-to-low) distribution, consistent with population distribution and corresponding anthropogenic emissions (Upadhyay et al., 2020; Dey et al., 2020).

Figure 2 show the spatial patterns of seasonal $PM_{2.5}$ and PM_{10} anomalies. The highest PM levels occurred in winter, especially in IGP (positive anomaly > 20 μg m^{-3} relative to the annual mean during 1980-2022). This enhancement is related to additional anthropogenic emissions (from space and water heating of households especially in cold places like IGP) and stable meteorological conditions (low boundary layer height and low wind speed) (Pandey et al., 2014; Tiwari et al., 2013). During the pre-monsoon (March-April-May), favorable meteorological conditions (increased boundary layer height due to increased temperature and wind speeds) reduce $PM_{2.5}$ concentrations in the IGP area(Dey et al., 2020). During the monsoon season (June to September), rainfall enhances PM deposition, resulting in a substantial reduction of PM concentrations. With the end of the monsoon (post-monsoon, October and November), less rainfall, lower temperatures, extensive open biomass burning (for heating), and reduced boundary layer heights exacerbate PM pollution(Nagpure et al., 2015; Kumari et al., 2021).

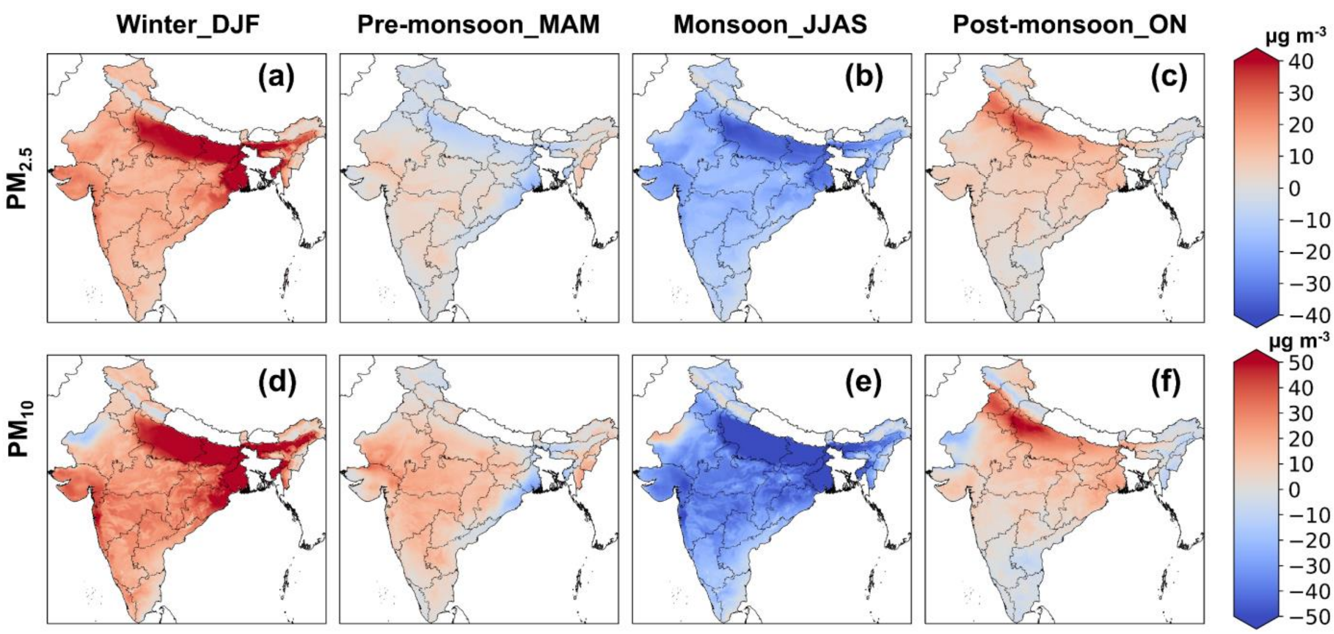


Figure 2: Spatial patterns of seasonal $PM_{2.5}$ and PM_{10} anomalies (the difference between seasonal mean and annual mean) in India during 1980-2022.

The long-term trends of aerosols in India can be better examined given the advantage of long temporal coverage of the LongPMInd dataset. The monthly $PM_{2.5}$ and PM_{10} anomalies from 1980 to 2022 in India and typical regions were firstly calculated (Fig. 3 and Fig. S6). PM concentrations slowly increased in India (0.19 μg m^{-3} $year^{-1}$) before 2000, the IGP and eastern India increased by 0.43 and 0.26 μg m^{-3} $year^{-1}$, respectively. With accelerated industrialization, anthropogenic emissions of primary particulate matter (PPM) and precursors of secondary aerosols (e.g., SO_2, NO, and NH_3) have increased since 2000 (Pandey et al., 2014; Nagpure et al., 2015), leading to significant increases of PM concentrations in most regions ($p<0.05$), except for western India (Fig. 4 and Fig. S7). $PM_{2.5}$ increased by 0.50 and 0.46 μg m^{-3} per year in the IGP and eastern

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

India during 2000-2017. In early 2018, the Indian government launched the National Clean Air Program (NCAP). $PM_{2.5}$ concentrations have declined significantly in the IGP (1.63 μg m^{-3} $year^{-1}$), western India (1.22 μg m^{-3} $year^{-1}$), and southern India (0.52 μg m^{-3} $year^{-1}$). However, PM concentrations in east-central India showed an increasing trend (Fig. 4), which may be related to emissions from mining activities and related industries and thermal power plants (Upadhyay et al., 2020).

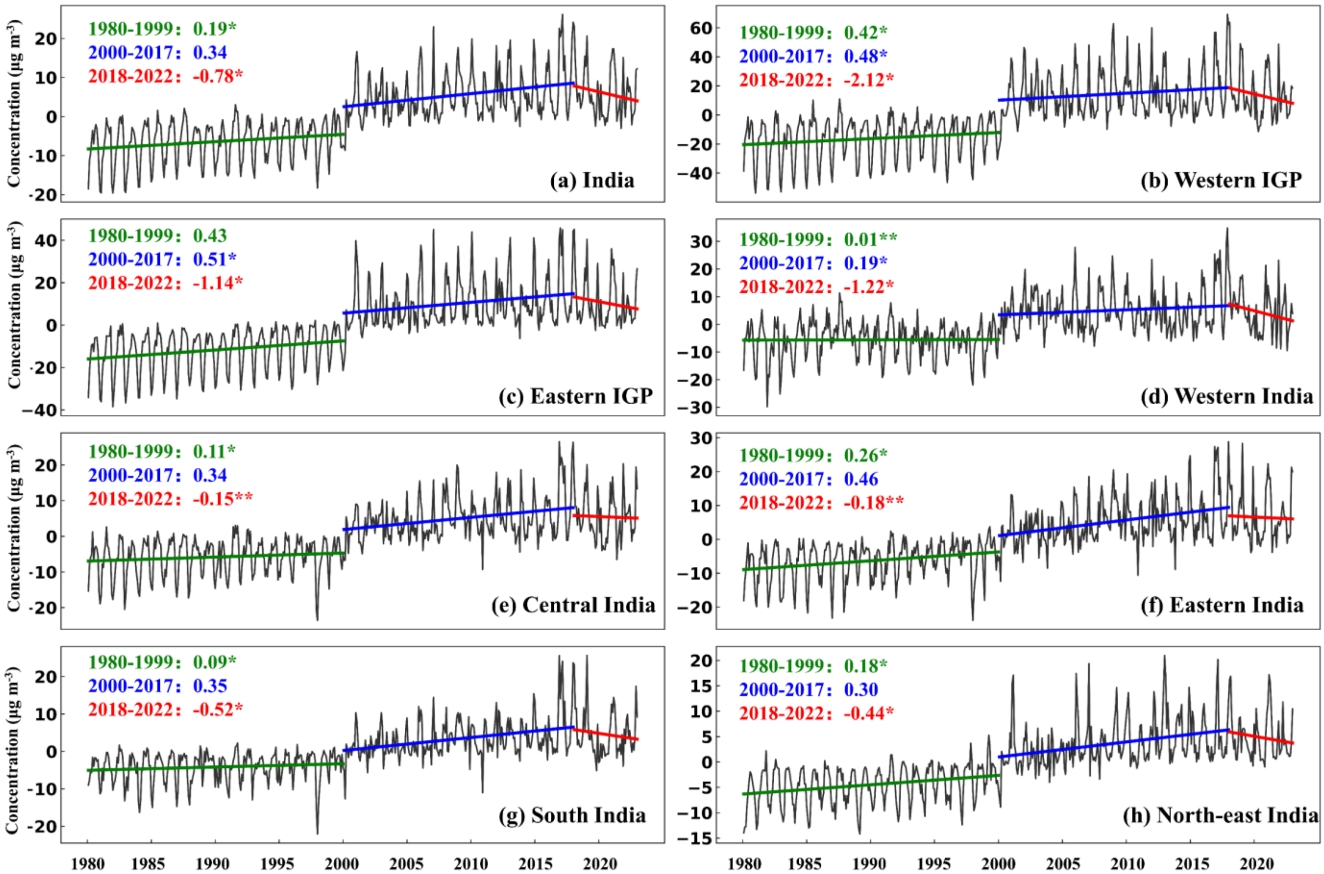


Figure 3: Time series of monthly $PM_{2.5}$ anomaly from 1980 to 2022 in India and typical regions. The colored straight lines are the linear regression trend (μg m^{-3} $year^{-1}$) for different periods in China, and * represent the significance of the trends (*mean p < 0.05 and ** mean p < 0.01).

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

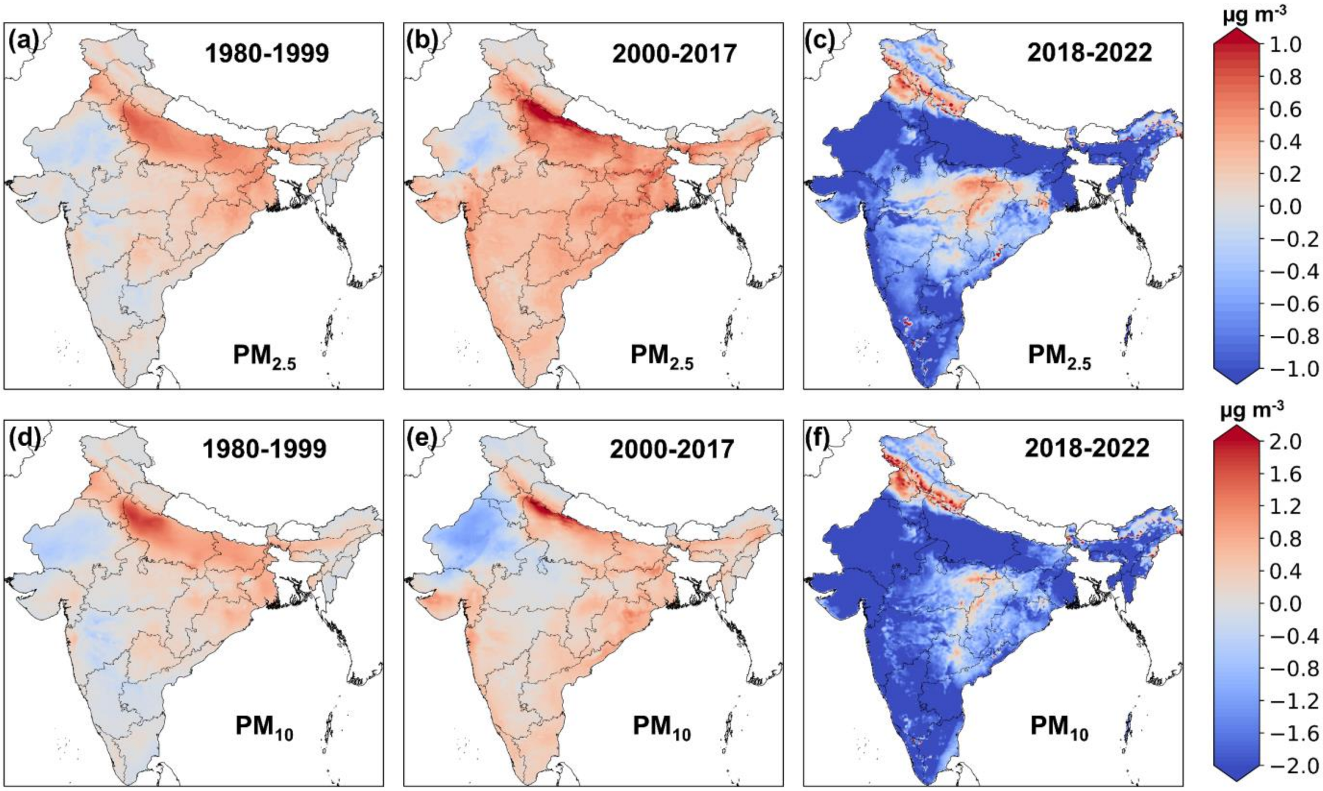


Figure 4: Spatial patterns of annual changes for $PM_{2.5}$ and PM_{10} (μg m^{-3} $year^{-1}$) during different periods (1980-1999, 2000-2017, and 2018-2022).

200 **3.4 Health burden analysis**

The health burden of $PM_{2.5}$ was estimated from 2000-2019 following the rapid increase in $PM_{2.5}$ concentrations after 2000. Using the database of GDB 2019, premature deaths attributed to $PM_{2.5}$ exposure were calculated for six diseases, including ischemic heart disease (CVD_IHD), chronic stroke (CVD_stroke), obstructive pulmonary disease (RESP_COPD), lung cancer (NEO_LUNG), lower respiratory infections (LRI), and diabetes mellitus type 2 (T2_DM) (Murray et al., 2020; Vos et al., 2020).

Figure 5 shows the changes of annual average $PM_{2.5}$ concentrations and corresponding attributed deaths, and Table S2 shows the uncertainties. $PM_{2.5}$ concentrations showed a fluctuating upward trend with a continuous increase of attributable premature mortality, from 0.73 (95 % Confidence Interval (CI): 0.65-0.80) million in 2000 to 1.22 (95 % CI: 1.03-1.41) million in 2019, with CVD_IHD, CVD_stroke, RESP_COPD, NEO_LUNG, LRI, and T2_DM caused an annual average of 0.35, 0.21, 0.21, 0.02, 0.12, 0.04 million premature mortality, respectively. $PM_{2.5}$-attributable deaths were counted by region (Fig. 5). The IGP had the highest attributable premature deaths, increasing from 0.36 million in 2000 to 0.60 million in 2019, due to high population density coupled with severe haze pollution (Dey et al., 2020; Pandey et al., 2021).

Open Access Earth System Science Data Discussions

To reduce premature deaths from $PM_{2.5}$ exposure, policies to mitigate $PM_{2.5}$ pollution should be implemented. In addition, appropriate health advice and enhanced medical facilities to reduce baseline mortality are also important to reduce the health burden (Maji et al., 2023). India has experienced rapid urbanization and large-scale population migration, which introduces uncertainty in health risk estimates for $PM_{2.5}$ (Shi et al., 2020). Country-level baseline disease rates were used, so regional differences were not accounted for due to lack of data, which could introduce some error. In addition, uncertainties in relative risk, population, and $PM_{2.5}$ concentrations may also introduce errors in health risk estimates.

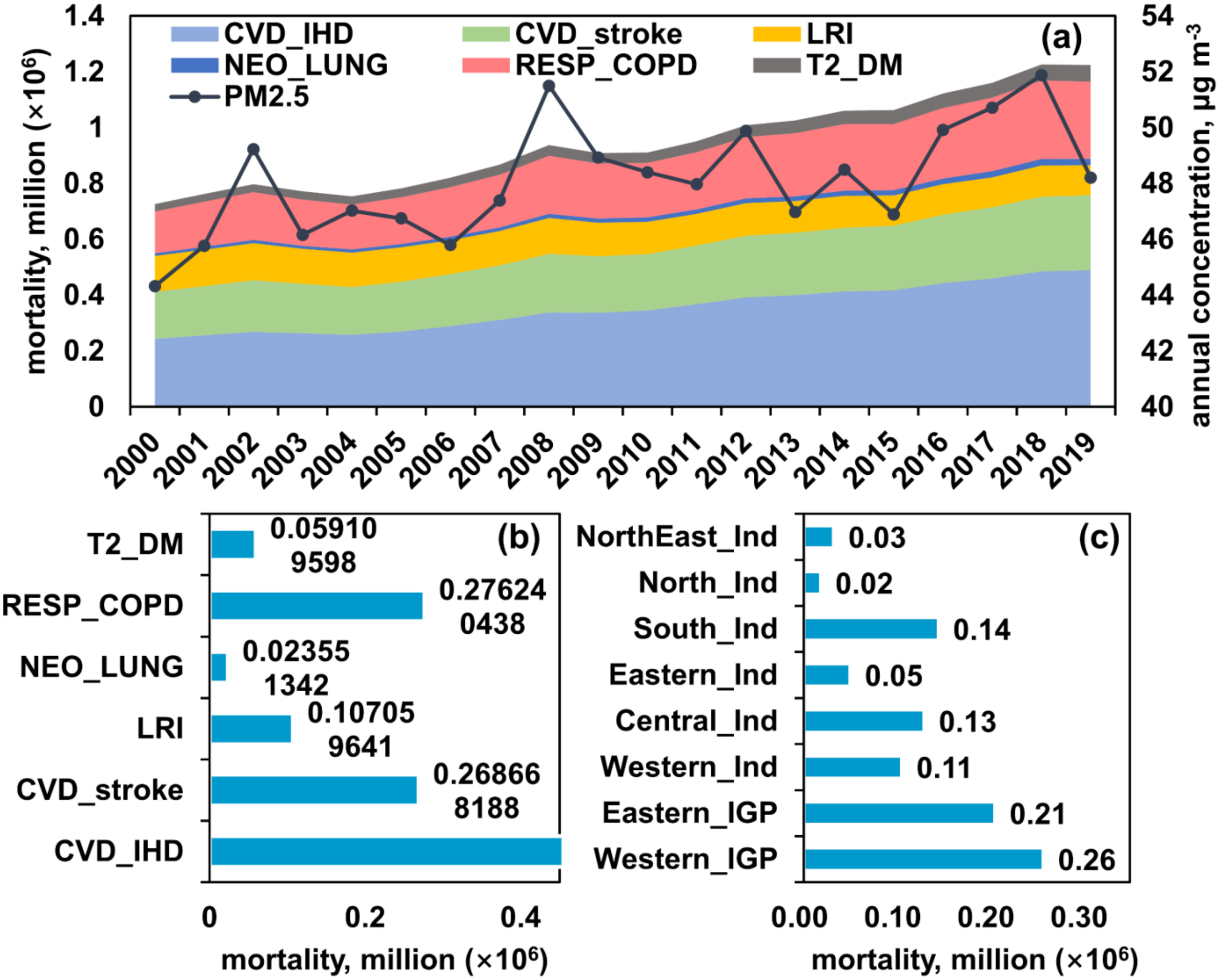
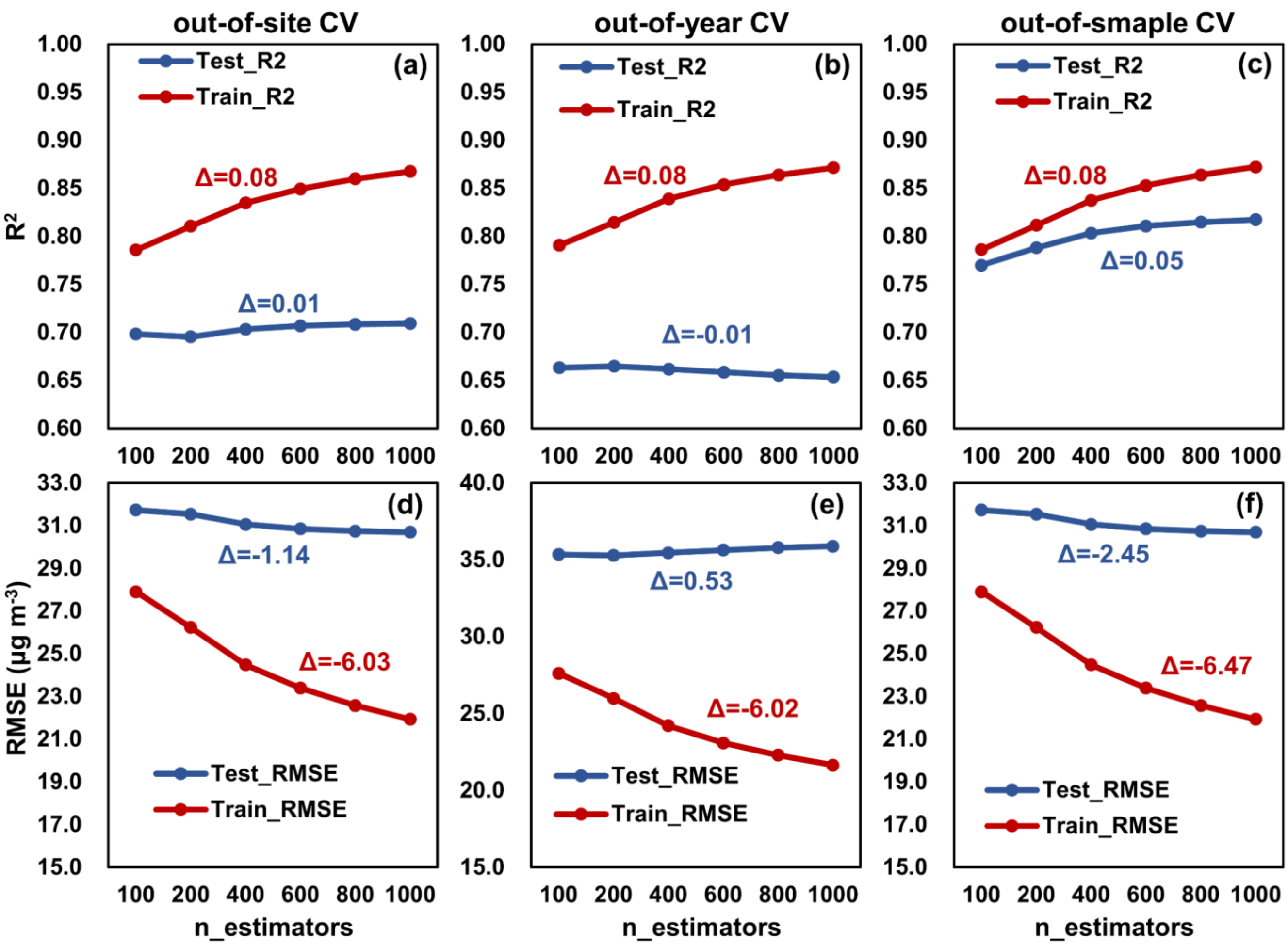


Figure 5: Annual mortalities due to $PM_{2.5}$-induced diseases in India during 2000-2019, including ischemic heart disease (CVD_IHD), chronic stroke (CVD_stroke), obstructive pulmonary disease (RESP_COPD), lung cancer (NEO_LUNG), lower respiratory infections (LRI), and diabetes mellitus type 2 (T2_DM). Subfigures b and c show statistical results for causes and regions.

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

3.5 Model complexity

Model complexity can be measured by the number of parameters the model has. As model complexity increases, the model is more capable to learn complex patterns in the data, but at the same time, it may lead to overfitting and inaccurate predictions of new and unseen data (Hu et al., 2021). The impact of the complexity of the tree-based LightGBM model on the performance of training and testing is analyzed. The number of trees (n_estimators) was used as a complexity proxy and the other hyperparameters were kept consistent. All three cross-validation results show that the increase of model complexity improves the model's fitting ability, increasing R^2 and decreasing RMSE. However, the increase in complexity did not improve the model's predictive performance. With n_estimators increasing from 100 to 1000, there was no significant change in R^2 for the out-of-site and out-of-year CV (-0.01 - 0.01), and the RMSE for the out-of-year CV on the contrary increased by 0.53. Out-of-sample CV showed an improvement in R^2 but with limited reduction in RMSE (-2.45). So, using only out-of-sample CV to select hyperparameters and evaluate the model is limiting, and out-of-site and out-of-year CV allows a more objective evaluation of the model's generalization ability.



https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

Figure 6: Three CV results of model complexity test for $PM_{2.5}$ estimation. The n_estimators is the number of trees, representing the complexity of LightGBM. Δ is the difference between the metrics with n_estimators = 1000 and n_estimators = 100. Units of RMSE and MAE are μg m^{-3}.

3.6 Uncertainties

Uncertainty in this study comes from two main sources: the machine learning model and the dataset used. Firstly, machine learning is essentially based on probability theory and is influenced by the distribution pattern of the target variable (PM1 concentration) (Breiman, 2001; Yang et al., 2021b). Due to the low frequency of extreme pollution scenarios, the model suffers from the problem of smoothing predictions, i.e. underestimating high pollution scenarios (Wei et al., 2021; Yu et al., 2023; Geng et al., 2021). In addition, machine learning has limitations in describing atmospheric physical and chemical processes, and it is difficult to fit complex, logistically long processes, such as secondary aerosol generation (Stirnberg et al., 2021; Li et al., 2023). Attempts have been made to incorporate physical constraints into neural networks to improve interpretability, but this approach is limited to spatially continuous two-dimensional data (Geiss et al., 2022). Other studies have shown that chemical reaction processes can be described by neural networks, it is still a challenge to efficiently couple them with CTMs (Huang et al., 2022; Huang and Seinfeld, 2022).

The second aspect is the uncertainties caused by the datasets. First, the label (observations) and corresponding features (MERRA2 and ERA5) has a long-tailed distribution with few high pollution records, so there is an issue of imbalance regression (Yang et al., 2021a). The model was trained with a bias towards denser observations, leading to underestimation of high pollution scenarios. For the problem of imbalanced regression there are currently main data-based solutions and model-based solutions (Ren et al., 2022a). Data-based solutions require acquiring more data or changing the data distribution by resampling. Model-based solutions increase the weighting of fewer samples (high pollution scenarios) by modifying the loss function. Both methods can improve the accuracy of fewer samples, but they are not suitable for the task of this study because the distribution of the data was altered. Therefore, more observations should be collected in the future to increase observations recorded for high pollution scenarios and mitigate the problem of imbalanced regression. In addition, observational data can only be collected for recent years (2018-2022), which may lead to uncertainties when inference PM concentrations for historical years. Out-of-year validation have been made to evaluate the model's predict ability for unobserved years, but changes in climate and human activities over the decades may affect the relationship among emissions, meteorology and PM concentrations, resulting in extra uncertainty.

Secondly, the uncertainty of the input feature sets (ERA5 and MERRA2) also affects the estimation results. The uncertainty of ERA5, a widely used meteorological reanalysis dataset, has been systematically analyzed. ERA5 has good accuracy for most meteorological factors, exceeding other reanalysis data (Muñoz-Sabater et al., 2021; Hersbach et al., 2020). With MODIS data as a reference, global mean surface temperature of ERA5-Land shows lower uncertainty (Muñoz-Sabater et al., 2021). For precipitation, ERA5 shows 77% correlation with monthly-mean Global Precipitation Climatology Project (GPCP) data (Hersbach et al., 2020). Compared to the pre-assimilation data, ERA5-land provides an improved fit to tropospheric winds and humidity (Hersbach et al., 2020).

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

MERRA2 is a global air pollution reanalysis dataset, published and maintained by NASA, which has been widely used for PM pollution studies in the Indian region, and its reliability has been extensively analyzed (Gueymard and Yang, 2020; Navinya et al., 2020; Buchard et al., 2017). For MERRA2-AOD, evaluation using AERONET observations showed that MERRA-2 outperformed the Copernicus Atmosphere Monitoring Service (CAMS) in most regions (Gueymard and Yang, 2020). Kumar et al. (2023) predicted ground-level $PM_{2.5}$ concentrations in India using only MERRA2 and machine learning methods, proving the reliability of MERRA2 data. In addition, prior to 2000, there was no assimilated satellite data for MERRA-2, which may be detrimental to the accuracy of the LongPMInd dataset, but the models trained in this study relied heavily on ERA5 (64 % relative contribution), with a minor contribution from MERRA2 (36 %).

Data availability

The LongPMInd dataset, including daily $PM_{2.5}$ and PM_{10} concentration (10km) for India during 1980-2022 is publicly accessible. All data are provided in NetCDF format and can be downloaded at https://zenodo.org/records/10073944 (Wang et al., 2023a).

Supporting Information

Research domain, feature importance, spatial and temporal patterns of $PM_{2.5}$ and PM_{10}, and uncertainty of estimated annual mortalities.

Author contribution

Shuai Wang: Methodology, Software, Writing - original draft. **Mengyuan Zhang**: Visualization, Validation. **Hui Zhao**: Data curation, Methodology. **Peng Wang**: Methodology, Writing - reviewing and editing. **Sri Harsha Kota**: Data curation. **Qingyan Fu**: Writing - reviewing and editing. **Cong Liu**: reviewing and editing. **Hongliang Zhang**: Conceptualization, Funding acquisition, Supervision, Writing - reviewing and editing.

Competing interests

The contact author has declared that neither they nor their co-authors have any competing interests.

Acknowledgment

This work was supported by the National Key R&D Program of China (2022YFC3701105), Co-fund DFG-NSFC Sino-German AirChanges project (448720203), National Natural Science Foundation of China (42077194/42061134008), and Shanghai International Science and Technology Partnership Project (No. 21230780200).

References

Miller, B. B. and Carter, C.: The test article, J. Sci. Res., 12, 135–147, doi:10.1234/56789, 2015.

Smith, A. A., Carter, C., and Miller, B. B.: More test articles, J. Adv. Res., 35, 13–28, doi:10.2345/67890, 2014.

Bai, K., Li, K., Ma, M., Li, K., Li, Z., Guo, J., Chang, N. B., Tan, Z., and Han, D.: LGHAP: the Long-term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion, Earth Syst. Sci. Data, 14, 907-927, 10.5194/essd-14-907-2022, 2022.

Bali, K., Dey, S., and Ganguly, D.: Diurnal patterns in ambient PM2.5 exposure over India using MERRA-2 reanalysis data, Atmos. Environ., 248, 118180, https://doi.org/10.1016/j.atmosenv.2020.118180, 2021.

Bali, K., Dey, S., Ganguly, D., and Smith, K. R.: Space-time variability of ambient PM2.5 diurnal pattern over India from 18-years (2000–2017) of MERRA-2 reanalysis data, Atmos. Chem. Phys. Discuss., 2019, 1-23, 10.5194/acp-2019-731, 2019.

Brauer, M., Guttikunda, S. K., K A, N., Dey, S., Tripathi, S. N., Weagle, C., and Martin, R. V.: Examination of monitoring approaches for ambient air pollution: A case study for India, Atmos. Environ., 216, 116940, https://doi.org/10.1016/j.atmosenv.2019.116940, 2019.

Breiman, L.: Random forests, Mach. Learn., 45, 5-32, 2001.

Buchard, V., Randles, C. A., da Silva, A. M., Darmenov, A., Colarco, P. R., Govindaraju, R., Ferrare, R., Hair, J., Beyersdorf, A. J., Ziemba, L. D., and Yu, H.: The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies, Journal of Climate, 30, 6851-6872, https://doi.org/10.1175/JCLI-D-16-0613.1, 2017.

Chowdhury, S., Dey, S., Di Girolamo, L., Smith, K. R., Pillarisetti, A., and Lyapustin, A.: Tracking ambient PM2. 5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset, Atmos. Environ., 204, 142-150, 2019.

Dandona, L., Dandona, R., Kumar, G. A., Shukla, D., Paul, V. K., Balakrishnan, K., Prabhakaran, D., Tandon, N., Salvi, S., and Dash, A.: Nations within a nation: variations in epidemiological transition across the states of India, 1990–2016 in the Global Burden of Disease Study, Lancet, 390, 2437-2460, 2017.

Dey, S., Purohit, B., Balyan, P., Dixit, K., Bali, K., Kumar, A., Imam, F., Chowdhury, S., Ganguly, D., Gargava, P., and Shukla, V. K.: A Satellite-Based High-Resolution (1-km) Ambient PM2.5 Database for India over Two Decades (2000–2019): Applications for Air Quality Management, Remote Sens., 12, 3872, 2020.

Dhandapani, A., Iqbal, J., and Kumar, R. N.: Application of machine learning (individual vs stacking) models on MERRA-2 data to predict surface PM2.5 concentrations over India, Chemosphere, 340, 139966, https://doi.org/10.1016/j.chemosphere.2023.139966, 2023.

Geiss, A., Silva, S. J., and Hardin, J. C.: Downscaling atmospheric chemistry simulations with physically consistent deep learning, Geosci. Model Dev., 15, 6677-6694, 2022.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., and Peng, Y.: Tracking Air Pollution in China: Near Real-Time PM2. 5 Retrievals from Multisource Data Fusion, Environ. Sci. Technol., 55, 12106-12115, 2021.

Grinsztajn, L., Oyallon, E., and Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data?, arXiv preprint arXiv:2207.08815, 2022.

Gueymard, C. A. and Yang, D.: Worldwide validation of CAMS and MERRA-2 reanalysis aerosol optical depth products using 15 years of AERONET observations, Atmos. Environ., 225, 117216, https://doi.org/10.1016/j.atmosenv.2019.117216, 2020.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Hu, X., Chu, L., Pei, J., Liu, W., and Bian, J.: Model complexity of deep learning: a survey, Knowledge and Information Systems, 63, 2585-2619, 10.1007/s10115-021-01605-0, 2021.

Huang, C., Hu, J., Xue, T., Xu, H., and Wang, M.: High-Resolution Spatiotemporal Modeling for Ambient PM2. 5 Exposure Assessment in China from 2013 to 2019, Environ. Sci. Technol., 55, 2152-2162, 2021.

Huang, J., Zhou, Y., and Yong, W.-A.: Data-driven discovery of multiscale chemical reactions governed by the law of mass action, Journal of Computational Physics, 448, 110743, 2022.

Huang, Y. and Seinfeld, J. H.: A neural network-assisted Euler integrator for stiff kinetics in atmospheric chemistry, Environ. Sci. Technol., 56, 4676-4685, 2022.

Jabbar, H. and Khan, R. Z.: Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study), Computer Science, Communication and Instrumentation Devices, 70, 978-981, 2015.

Katoch, V., Kumar, A., Imam, F., Sarkar, D., Knibbs, L. D., Liu, Y., Ganguly, D., and Dey, S.: Addressing Biases in Ambient PM2.5 Exposure and Associated Health Burden Estimates by Filling Satellite AOD Retrieval Gaps over India, Environ. Sci. Technol., 57, 19190-19201, 10.1021/acs.est.3c03355, 2023.

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., Ye, Q. W., and Liu, T. Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, Dec 04-09, WOS:000452649403021, 2017.

Kumar, V., Malyan, V., Sahu, M., Biswal, B., Pawar, M., and Dev, I.: Spatiotemporal analysis of fine particulate matter for India (1980–2021) from MERRA-2 using ensemble machine learning, Atmospheric Pollution Research, 14, 101834, https://doi.org/10.1016/j.apr.2023.101834, 2023.

Kumari, S., Verma, N., Lakhani, A., and Kumari, K. M.: Severe haze events in the Indo-Gangetic Plain during post-monsoon: Synergetic effect of synoptic meteorology and crop residue burning emission, Science of the Total Environment, 768, 145479, 2021.

Li, T., Zhang, Q., Peng, Y., Guan, X., Li, L., Mu, J., Wang, X., Yin, X., and Wang, Q.: Contributions of Various Driving Factors to Air Pollution Events: Interpretability Analysis from Machine Learning Perspective, Environ. Int., 107861, 2023.

Maheshwarkar, P., Ralhan, A., Sunder Raman, R., Tibrewal, K., Venkataraman, C., Dhandapani, A., Kumar, R. N., Mukherjee, S., Chatterje, A., Rabha, S., Saikia, B. K., Bhardwaj, A., Chaudhary, P., Sinha, B., Lokhande, P., Phuleria, H. C., Roy, S., Imran, M., Habib, G., Azharuddin Hashmi, M., Qureshi, A., Qadri, A. M., Gupta, T., Lian, Y., Pandithurai, G., Prasad, L., Murthy, S., Deswal, M., Laura, J. S., Chhangani, A. K., Najar, T. A., and Jehangir, A.: Understanding the Influence of Meteorology and Emission Sources on PM2.5 Mass Concentrations Across India: First Results From the COALESCE Network, J. Geophys. Res.-Atmos., 127, e2021JD035663, https://doi.org/10.1029/2021JD035663, 2022.

Maji, K. J., Namdeo, A., and Bramwell, L.: Driving factors behind the continuous increase of long-term PM2.5-attributable health burden in India using the high-resolution global datasets from 2001 to 2020, Science of The Total Environment, 866, 161435, https://doi.org/10.1016/j.scitotenv.2023.161435, 2023.

Martin, R. V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., and Dey, S.: No one knows which city has the highest concentration of fine particulate matter, Atmospheric Environment: X, 3, 100040, 2019.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349-4383, 10.5194/essd-13-4349-2021, 2021.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., and Abdollahpour, I. J. T. L.: Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet, 396, 1223-1249, 2020.

Nagpure, A. S., Ramaswami, A., and Russell, A.: Characterizing the spatial and temporal patterns of open burning of municipal solid waste (MSW) in Indian cities, Environ. Sci. Technol., 49, 12904-12912, 2015.

Navinya, C. D., Vinoj, V., and Pandey, S. K.: Evaluation of PM2.5 Surface Concentrations Simulated by NASA's MERRA Version 2 Aerosol Reanalysis over India and its Relation to the Air Quality Index, Aerosol and Air Quality Research, 20, 1329-1339, 10.4209/aaqr.2019.12.0615, 2020.

Pandey, A., Sadavarte, P., Rao, A. B., and Venkataraman, C.: Trends in multi-pollutant emissions from a technology-linked inventory for India: II. Residential, agricultural and informal industry sectors, Atmos. Environ., 99, 341-352, https://doi.org/10.1016/j.atmosenv.2014.09.080, 2014.

Pandey, A., Brauer, M., Cropper, M. L., Balakrishnan, K., Mathur, P., Dey, S., Turkgulu, B., Kumar, G. A., Khare, M., and Beig, G.: Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019, The Lancet Planetary Health, 5, e25-e38, 2021.

Pant, P., Lal, R. M., Guttikunda, S. K., Russell, A. G., Nagpure, A. S., Ramaswami, A., and Peltier, R. E.: Monitoring particulate matter in India: recent trends and future outlook, Air Quality, Atmosphere & Health, 12, 45-58, 2019.

Ren, J., Zhang, M., Yu, C., and Liu, Z.: Balanced mse for imbalanced visual regression, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7926-7935,

Ren, X., Mi, Z., Cai, T., Nolte, C. G., and Georgopoulos, P. G.: Flexible Bayesian Ensemble Machine Learning Framework for Predicting Local Ozone Concentrations, Environ. Sci. Technol., 56, 3871-3883, 10.1021/acs.est.1c04076, 2022b.

Sayeed, A., Lin, P., Gupta, P., Tran, N. N. M., Buchard, V., and Christopher, S.: Hourly and Daily PM2.5 Estimations Using MERRA-2: A Machine Learning Approach, Earth Space Sci., 9, e2022EA002375, https://doi.org/10.1029/2022EA002375, 2022.

Shi, G., Lu, X., Deng, Y., Urpelainen, J., Liu, L.-C., Zhang, Z., Wei, W., and Wang, H.: Air pollutant emissions induced by population migration in China, Environ. Sci. Technol., 54, 6308-6318, 2020.

Stirnberg, R., Cermak, J., Kotthaus, S., Haeffelin, M., Andersen, H., Fuchs, J., Kim, M., Petit, J. E., and Favez, O.: Meteorology-driven variability of air pollution (PM1) revealed with explainable machine learning, Atmos. Chem. Phys., 21, 3919-3948, 10.5194/acp-21-3919-2021, 2021.

Tiwari, S., Srivastava, A. K., Bisht, D. S., Parmita, P., Srivastava, M. K., and Attri, S.: Diurnal and seasonal variations of black carbon and PM2. 5 over New Delhi, India: Influence of meteorology, Atmospheric Research, 125, 50-62, 2013.

Upadhyay, A., Dey, S., and Goyal, P.: A comparative assessment of regional representativeness of EDGAR and ECLIPSE emission inventories for air quality studies in India, Atmos. Environ., 223, 117182, 2020.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., and Abdelalim, A.: Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet, 396, 1204-1222, 2020.

Wang, K. C., Dickinson, R. E., Wild, M., and Liang, S.: Atmospheric impacts on climatic variability of surface incident solar radiation, Atmos. Chem. Phys., 12, 9581-9592, 10.5194/acp-12-9581-2012, 2012.

Wang, S., Kota, S. H., and Zhang, H.: LongPMInd: long-term (1980-2022) daily ground particulate matter datasets in India, Zenodo [dataset], 10.5281/zenodo.10073944, 2023a.

Wang, S., Wang, P., Zhang, R., Meng, X., Kan, H., and Zhang, H.: Estimating particulate matter concentrations and meteorological contributions in China during 2000–2020, Chemosphere, 330, 138742, https://doi.org/10.1016/j.chemosphere.2023.138742, 2023b.

Wang, S., Zhang, M., Gao, Y., Wang, P., Fu, Q., and Zhang, H.: Diagnosing drivers of PM2.5 simulation biases from meteorology, chemical composition, and emission sources using an efficient machine learning method, EGUsphere, 2023, 1-14, 10.5194/egusphere-2023-1531, 2023c.

Wang, S., Wang, P., Qi, Q., Wang, S., Meng, X., Kan, H., Zhu, S., and Zhang, H.: Improved estimation of particulate matter in China based on multisource data fusion, Science of The Total Environment, 161552, 2023d.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sensing of Environment, 252, 112136, 2021.

Xue, T., Zheng, Y., Geng, G., Xiao, Q., Meng, X., Wang, M., Li, X., Wu, N., Zhang, Q., and Zhu, T.: Estimating Spatiotemporal Variation in Ambient Ozone Exposure during 2013–2017 Using a Data-Fusion Model, Environ. Sci. Technol., 54, 14877-14888, 10.1021/acs.est.0c03098, 2020.

Yang, X., Zhao, C. F., Zhou, L. J., Wang, Y., and Liu, X. H.: Distinct impact of different types of aerosols on surface solar radiation in China, J. Geophys. Res.-Atmos., 121, 6459-6471, 10.1002/2016jd024938, 2016.

https://doi.org/10.5194/essd-2024-34
Preprint. Discussion started: 6 February 2024

Yang, Y., Zha, K., Chen, Y., Wang, H., and Katabi, D.: Delving into deep imbalanced regression, International Conference on Machine Learning, 11842-11851,

Yang, Y. Z., Zha, K. W., Chen, Y. C., Wang, H., and Katabi, D.: Delving into Deep Imbalanced Regression, International Conference on Machine Learning (ICML), Electr Network, Jul 18-24, WOS:000768182701085, 2021.

Ying, X.: An overview of overfitting and its solutions, Journal of physics: Conference series, 022022,

Yu, W., Ye, T., Zhang, Y., Xu, R., Lei, Y., Chen, Z., Yang, Z., Zhang, Y., Song, J., and Yue, X.: Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study, The Lancet Planetary Health, 7, e209-e218, 2023.

Zhang, T. N., He, W. H., Zheng, H., Cui, Y. P., Song, H. Q., and Fu, S. L.: Satellite-based ground PM2.5 estimation using a gradient boosting decision tree, Chemosphere, 268, 10.1016/j.chemosphere.2020.128801, 2021.