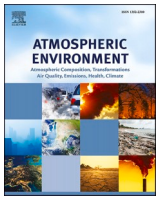
Atmospheric Environment 338 (2024) 120834

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv



Extracting regional and temporal features to improve machine learning for hourly air pollutants in urban India

Shuai Wang [a], Mengyuan Zhang [a], Hui Zhao [b], Peng Wang [c,d], Sri Harsha Kota [e], Qingyan Fu [f], Hongliang Zhang [a,d,g,*]

[a] Department of Environmental Science and Engineering, Fudan University, Shanghai, 200438, China
[b] School of Resources and Environmental Engineering, Jiangsu University of Technology, Changzhou, 213001, China
[c] Department of Atmospheric and Oceanic Sciences, and Institute of Atmospheric Sciences, Fudan University, Shanghai, 200438, China
[d] IRDR ICoE on Risk Interconnectivity and Governance on Weather/Climate Extremes Impact and Public Health, Fudan University, Shanghai, China
[e] Department of Civil Engineering, Indian Institute of Technology, Delhi, 110016, India
[f] Shanghai Environmental Monitoring Center, Shanghai, 200235, China
[g] Institute of Eco-Chongming (IEC), Shanghai 200062, China

HIGHLIGHTS

- Regional and temporal features were extracted to improve $PM_{2.5}$ estimation.
- Feature extraction reduced the RMSE of $PM_{2.5}$ and PM_{10} estimation by 21% and 19%.
- Model explanation reveals a strong impact of boundary layer height on winter $PM_{2.5}$.

ARTICLE INFO

Keywords:
Machine learning
Feature extraction
$PM_{2.5}$ datasets
Air pollution
India

ABSTRACT

India is suffering from severe particulate matter (PM, including $PM_{2.5}$ and PM_{10}) pollution, while limited ground observations are insufficient to support a comprehensive understanding of its health risks. Machine learning (ML) has the potential to improve the estimation of PM distribution and exposure efficiently. Regional transport as well as accumulation and dispersion processes of PM and its components, which have significant impacts on PM concentrations, are crucial when building ML models, especially for sparsely observed regions like India. Here, geographic and temporal-rolling weighting methods were used to separately extract regional and temporal features for improving the performance of the ML model. The incorporation of temporal and regional features into the ML model significantly improved ML model performance, with root mean square error (RMSE) reduced by 21 % and 19% for $PM_{2.5}$ and PM_{10} estimation, as well as an improvement in model underestimation for the heavy pollution scenarios. The spatial-temporal model shows out-of-sample test CV coefficients of determination (R^2) of 0.87 and 0.88 for hourly $PM_{2.5}$ and PM_{10}. The ML model predicts an annual nationwide concentration of 68.3 μg/m^3 for $PM_{2.5}$ with a north (high, especially in Indo-Gangetic Plain) to south (low) distribution, which is consistent with high satellite aerosol optical depth (AOD) values. Boundary layer height is identified as the main meteorological factor influencing $PM_{2.5}$ concentrations in winter. Characterizing the regional transport and cumulative dispersion processes of pollutants by extracting features can help in machine learning training, and this method can be further improved and applied to other studies.

1. Introduction

India has been exposed to severe air pollution in recent years, especially for particulate matter (PM, including $PM_{2.5}$ with diameters <2.5 μm and PM_{10} with diameters <10 μm), a complex mixture influenced by meteorology, primary emissions, and atmospheric chemical processes (Dey et al., 2020; Maheshwarkar et al., 2022; Organization, 2021). $PM_{2.5}$ is one of the leading causes of health burden in India, with

* Corresponding author. Department of Environmental Science and Engineering, Fudan University, Shanghai, 200438, China.
E-mail address: zhanghl@fudan.edu.cn (H. Zhang).

https://doi.org/10.1016/j.atmosenv.2024.120834
Received 8 July 2024; Received in revised form 12 September 2024; Accepted 17 September 2024
Available online 18 September 2024

adverse effects on ecosystems and climate (de Bont et al., 2024; Hammer et al., 2020; Murray et al., 2020; Wang et al., 2012). The Central Pollution Control Board (CPCB) has established and maintains ground-based monitoring network in India. However, these monitoring sites are unevenly distributed and mainly located in urban areas (residential and industrial areas), with limited site numbers (monitoring density: 0.6 stations/million people) (Brauer et al., 2019). These limitations make it difficult for ground-based observations to support air quality management on a regional scale. Chemical transport models (CTMs), such as The Community Multiscale Air Quality (CMAQ) model, are widely used as a useful tool for air pollution modeling, while due to uncertainties in emission inventories and meteorological fields, as well as simplified chemical mechanisms, CTMs suffer from significant biases (Hu et al., 2014, 2016).

With the rapid development of artificial intelligence, machine learning (ML) has been widely used in air pollution modeling due to its capability to describe complex nonlinear relationships (Wang et al., 2023b, 2023c; Wei et al., 2021b). Previous studies have shown that for tabular data (e.g., observation datasets of air pollutants), tree-based models usually outperform deep neural networks due to the limited amount of data and features as well as relative sample latent information, and are therefore widely developed and applied (Grinsztajn et al., 2022). Modeling air pollution using ML methods with only local data has limitations in describing the regional effects of meteorology and emissions (e.g., transport by wind), especially for sparsely observed regions like India (Dey et al., 2020; Qiu et al., 2022; Wang et al., 2023c). Wei et al. (2021a) encoded latitude and longitude to indirectly describe the spatial information, but this method can hardly interpret the contribution of the regional processes. Qiu et al. (2022) added unprocessed regional meteorological features (meteorological variable values of all grids around the observation site) to build ML models, while the excessive number of features led to dimensional explosion and overfitting risk.

Deep neural networks, such as convolutional neural nets, have the potential to extract spatial information, but most pre-trained image feature extraction networks are used for tasks such as target detection and segmentation which do not apply to meteorological, emission and air pollution maps (He et al., 2016; Ping Tian, 2013; Zhao and Du, 2016). Retraining of feature extraction networks applicable to air pollution images relies on large GPU computational resources and large-scale, high-quality datasets, and is therefore difficult to achieve. In addition, the traditional image feature extraction network outputs high-dimensionality features (Ping Tian, 2013). For example, Pre-trained VGG16 (Simonyan and Zisserman, 2014), as a common algorithm for image feature extraction, outputs features with 4096 dimensions, which is hard to interpret. Recent study have proposed causally inferred representation learning methods to model air pollution, using equilibrium fraction theory to learn representations of nonlocal information as scalars or vectors defined for each observation unit (Tec et al., 2023). The high demand for GPU computational resources and complex deployment limits its wide application.

Geographically weighted regression (GWR) can quantify spatial heterogeneity and describe spatial relationships (Fotheringham et al., 2015; Fotheringham and Oshan, 2016). Previous studies have experimented to fit particulate matter pollution using GWR and showed poor prediction accuracy, due to its weaker ability to characterize complex nonlinear relationships compared to ML methods (Ma et al., 2014). However, the geographically weighted approach can describe the spatial correlation and spatial heterogeneity and therefore has the potential to extract spatial features to characterize the regional effects of meteorology and emissions in air pollution modeling. Historical information is also crucial to describe the accumulation and dissipation processes of pollutants. Wei et al. (2021a) coded temporal variables to provide historical information which improved model performance but had interpretability limitations. Jin et al. (2019) used features from a past period to describe temporal features, which substantially increases the data dimensionality when feature numbers increase and with finer temporal resolution. Therefore, rational extraction of meteorological and pollutant lag features has the potential to describe the accumulation and dissipation processes of pollutants while reducing the feature dimensionality.

In this study, a geographically weighted approach was used to extract regional features for meteorology and emissions, and a rolling weighted approach was used to characterize the accumulation and dissipation processes of pollutants. An efficient tree-based ML model lightGBM was used to predict hourly $PM_{2.5}$ and PM_{10} in India in 2018. The model performance was evaluated using two independent cross-validation (CV) methods. Model prediction results were explained with an interpretable ML approach.

2. Materials and methods

2.1. Data sources

Hourly $PM_{2.5}$ observations for 2018 in India were collected from the national ambient air quality monitoring network (Continuous Automatic Air Quality Monitoring Station: CAAQMS) of the Central Pollution Control Board (CPCB). The distribution of 73 monitoring stations is shown in Fig. S1. The stations are sparsely distributed, except in the region of Delhi. All stations are located in urban areas (Pant et al., 2019). The temporal resolution of the observed data is 1 h. More information can be found at https://airquality.cpcb.gov.in/ccr/. The observations were selected by excluding extreme values (<0.1 % quantile and >99.9 % quantile), duplicate values, negative values, and values where $PM_{2.5}$ exceeded PM_{10}. The observation data was divided into 10-km grids. For stations distributed on the same grids, mean values were calculated. The data records with discontinuous observations due to missing values (about 2%) were also excluded.

The fifth generation ECMWF atmospheric reanalysis datasets ERA5-Land were collected to provide meteorological fields, which have $0.1^\circ \times 0.1^\circ$ horizontal resolution and hourly temporal resolution. ERA5 hourly datasets on single levels with $0.25^\circ \times 0.25^\circ$ horizontal resolution were used to provide meteorological fields on the ocean and concatenated with ERA5-land. Emissions Database for Global Atmospheric Research (EDGAR v6.1) was used to provide anthropogenic emissions for the ML model, including primary particulates (PPM), sulfur dioxide (SO_2), ammonia (NH_3), nitrogen oxides (NOx), volatile organic compounds (VOCs). The biogenic emissions were generated by the Model for Emissions of Gases and Aerosols from Nature (MEGAN v2.1) and the open burning emissions were obtained from The Fire INventory from NCAR (FINN) (Guenther et al., 2006, 2012; Wiedinmyer et al., 2011).

2.2. Features extraction

Regional information of meteorology and emissions were extracted using a geo-weighting approach (Fig. 1). For each grid, the surrounding square windows were selected for inverse distance weighting, and the weighting algorithm is shown in Equations (1) and (2). The weighted results were used as features to be concatenated with the local features and fed into the ML model to make the final prediction. The influence of window size, weighting function, and bandwidth on ML modeling was tested by sensitivity experiments (Figs. S2 and S3). The results show that increasing the window size has a small effect on the model performance, larger spatially weighted windows can slightly improve the model performance but run the risk of overfitting, and output overly smoothed prediction. Guo et al. (2019) used a chemical transport model to track the regional transport of particulate matter in the Indian region and showed that $PM_{2.5}$ has a limited regional transmission distance. Therefore, a 200 km window size was selected for regional feature extraction. Several kernel functions were tested with little influence on model performance. So, the Gaussian kernel function, as a widely used algorithm, was selected for spatial weighting (Oshan et al., 2020; You et al.,

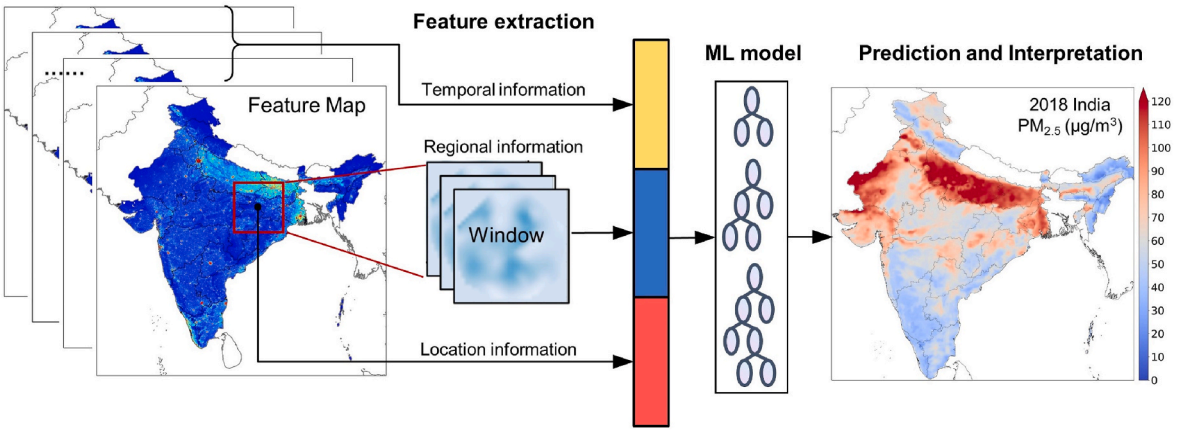


Fig. 1. Schematic of the regional and temporal features extraction in our study.

2016). Bandwidth was also tested which showed a significant impact on model performance. Here, the bandwidth was set to half the window size (100 km) to avoid over-smoothing in the spatial prediction.

$$\mathrm{des}(p_i, p_0) = \sqrt{\left(p_{i,x} - p_{0,x}\right)^2 + \left(p_{i,y} - p_{0,y}\right)^2} \quad (1)$$

$$\mathrm{w}(p_i, p_0) = \mathrm{kernal}\left(\frac{\mathrm{des}(p_i, p_0)}{bw}\right) \quad (2)$$

The des is Euclidean distance; p_i is the position of i grid; p_0 is the local grid; x is the number of grids in the X directions; y is the number of grids in the Y directions; bw is bandwidth; kernel is Gaussian weighting (Anselin and Rey, 2010); w (p_i, p_0) is the weight for grid p_i.

The historical temporal information of meteorology and emissions was extracted using a rolling weighting approach (Fig. 1). A period of historical time (e.g., 8 h) was selected for rolling weighting, and the results obtained were again combined with local features. The datasets were organized in a table where rows were the monitoring sites at different time points (around 324,000) and columns were local, geographically weighted, and rolling-weighted features, which were fed to the ML model to make predictions. The same sensitivity experiments were also designed to test the influence of window size, and weighting functions (Figs. S2 and S3). A 24-h window was used to extract historical information. Mean weighting slightly outperforms several other weighting methods and is therefore adopted.

2.3. Machine learning

The lightGBM was used to model $PM_{2.5}$ and PM_{10} in this study, which is an optimization of Gradient Boosted Decision Tree (GBDT) (Ke et al., 2017). Our previous study has proven its high accuracy, fast speed, and good robustness (Wang et al., 2023c). Several meteorological factors highly relevant to air pollutants generation, transportation, and deposition (Table S1) were selected for ML model building (Chen et al., 2020; Meng et al., 2019; Xiao et al., 2021). Emissions from anthropogenic, biogenic, and open burning were also included. The Gain is used to measure feature importance and select features, which is the degree of model improvement after splitting at a node during the tree growth. The top 20 features of meteorology and emission with the highest relative importance and their regional and temporal features were used for ML model training.

The out-of-sample grid search CV method was used to select the best hyperparameters and the optimal model. A hyperparameter selection algorithm (SI: Algorithm 1) was designed to ensure model robustness. Looping increases the model complexity (e.g., the number of trees), ending and returning the hyperparameters when there is no significant decrease in the model's testing RMSE (<0.01) or when the difference between the training RMSE and the predicted RMSE does not increase significantly (<0.05). In addition to out-of-sample CV, an out-of-site CV was used to assess model performance. The data set was randomly divided into 10 subsets by sites, sequentially taking one for testing and the rest 9 subsets for training, repeating 10 times, and the average is taken as the validation result. Sites were randomly divided 20 times, and the average was calculated as the validation result. This method can measure the spatial predictive power of the model (Geng et al., 2021; Wei et al., 2021a, 2023).

One-year hourly concentration of $PM_{2.5}$ and PM_{10} were estimated. Limited by computational resources, we do not estimate long-term hourly datasets. Our another study conducted daily-frequency particulate matter estimation for the last 40 years (Wang et al., 2024) and open-sourced the dataset at zenodo (Wang et al., 2023a), which provides a long-term data record for the Indian region.

2.4. Model explanation

The SHapley Additive exPlanation (SHAP) approach was used to interpret the ML model's prediction results (Lundberg and Lee, 2017). SHAP is a game theory-based approach using the classic Shapley value, which assigns an importance value to each feature for a particular prediction. The LightGBM model has a relatively simple structure compared to complex deep neural network with a small number of parameters, the decision tree only depends on the Gain of one feature per split, which makes the decision-making process of each node transparent to some extent, and thus the SHAP values can be directly utilized in the tree structure to compute the contribution of each feature to each prediction. The background dataset was set to 1000 random training samples. The interventional method was used to handle correlated input features (Janzing et al., 2020).

The meteorological normalization was used to decouple the effects of meteorology (eq (3) and eq (4)) (Hou et al., 2022; Wang et al., 2023c). For observations at a specific time, all meteorological features were randomly sampled 1000 times on a national scale to represent the average meteorological state, keeping the rest of the features unchanged (emission and temporal features), and 1000 predictions were generated using the trained ML model and averaged to compute the meteorological normalized $PM_{2.5}$ concentration at a specific time (the emission level under average meteorological conditions, $PM_{2.5,emiss}$). ML model for meteorological driven $PM_{2.5}$ ($PM_{2.5}$,meteo) were reconstructed. We retrained a model with the feature as meteorological variables and the target as $PM_{2.5,meteo}$, for the interpretation of the results. The contribution of each meteorological feature was quantified using the SHAP method.

$$PM_{2.5,emiss} = \frac{1}{1000}\sum_{i=1}^{1000} C_{i,pred} \quad (3)$$

$$PM_{2.5,meteo} = PM_{2.5,est} - PM_{2.5,emiss} \quad (4)$$

where $C_{i,pred}$ is the model-predicted PM concentration for meteorological conditions at a given time I; $PM_{2.5,est}$ is the estimated PM concentration; $PM_{2.5,emiss}$ is emission contribution; $PM_{2.5,meteo}$ is the meteorological contribution.

3. Results and discussion

3.1. Feature importance

Correlations between selected features and target air pollutants ($PM_{2.5}$ and PM_{10}) were first investigated (Fig. 2). $PM_{2.5}$ and PM_{10} show similar correlations with meteorological and emission-related features. Among the local features, except for SP (0.29), SSRD (0.20), and EVAP (0.14), the meteorological features are negatively correlated with $PM_{2.5}$ concentrations, especially TCLOUD (−0.54), DEWP2 (−0.51), and TPREC (−0.50). Conversely, most emission-related features are positively correlated with $PM_{2.5}$ concentrations, especially for primary particulate matter PPM (0.32), POC (0.36), and SO_2 (0.19). Among the regional characteristics, SO_2 and primary sulfate are significantly and positively correlated with $PM_{2.5}$, implying their contribution to $PM_{2.5}$ concentrations.

The relative importance of meteorological and emissions features is similar when predicting $PM_{2.5}$ as it is when predicting PM_{10}. Extracted temporal and spatial features show high importance (>30 %), implying the validity of the features. Surface pressure (SP) and wind speed (UWIND10 and VWIND10) show a large contribution among temporal, spatial, and local features. Compared to local information, temporal cloud cover and precipitation show larger contributions, indicating the influence of wet deposition on $PM_{2.5}$ concentrations in the last 8 h.

There is no duplicate information between temporal and local features, but there is duplicate information for regional and local features. The regional features are derived by weighting a 20*20 spatial window which contains a total of 400 data points and contains data points for the local features. Therefore, there is a 1/400 overlap between them. Covariance may exist between different features, but it does not affect the predictive ability of the tree model (Ke et al., 2017; Wang et al., 2024).

3.2. Model performance

Table 1 shows the training and testing out-of-sample CV results of $PM_{2.5}$ and PM_{10} for original, spatial, temporal, and spatio-temporal models, and Table S2 shows the corresponding out-of-site CV results. The original model was built with only local features, the spatial model was built with local and spatial features, the temporal model was built with local and temporal features, and the spatio-temporal model was built with local, spatial, and temporal features. Compared to previous studies, we achieved a lower RSME (Dhandapani et al., 2023). All four types of ML models outperformed CMAQ, and the spatial-temporal model of $PM_{2.5}$ showed 56% and 25% lower RMSE for out-of-sample and out-of-site CV than CMAQ. The spatio-temporal model shows higher out-of-sample CV accuracy than the original model, with test R^2, RMSE, and MAE improved by 10 %, 21 %, and 22 % for $PM_{2.5}$, and by

Table 1
Training and testing out-of-sample CV results of hourly $PM_{2.5}$ and PM_{10} for four types of ML models and CMAQ simulations in India. Original: model trained with only local information; Spatial training with local and spatial information; Temporal: training with local and temporal information; Spatio-temporal: training with local, spatial, and temporal information. 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
$PM_{2.5}$	Original	0.79	0.87	37.23	29.13	21.71	18.31
	Spatial	0.80	0.89	35.72	27.35	20.65	17.20
	Temporal	0.86	0.92	30.03	22.11	17.26	14.01
	Spatio-temporal	0.87	0.92	29.29	21.71	16.85	13.84
	CMAQ	–	–	68.11	–	40.61	–
PM_{10}	Original	0.81	0.89	63.23	48.87	40.66	33.41
	Spatial	0.82	0.90	61.62	46.79	39.39	31.94
	Temporal	0.87	0.93	51.14	38.21	33.26	26.15
	Spatio-temporal	0.88	0.94	51.27	37.12	32.40	25.56
	CMAQ	–	–	151.96	–	98.2	–

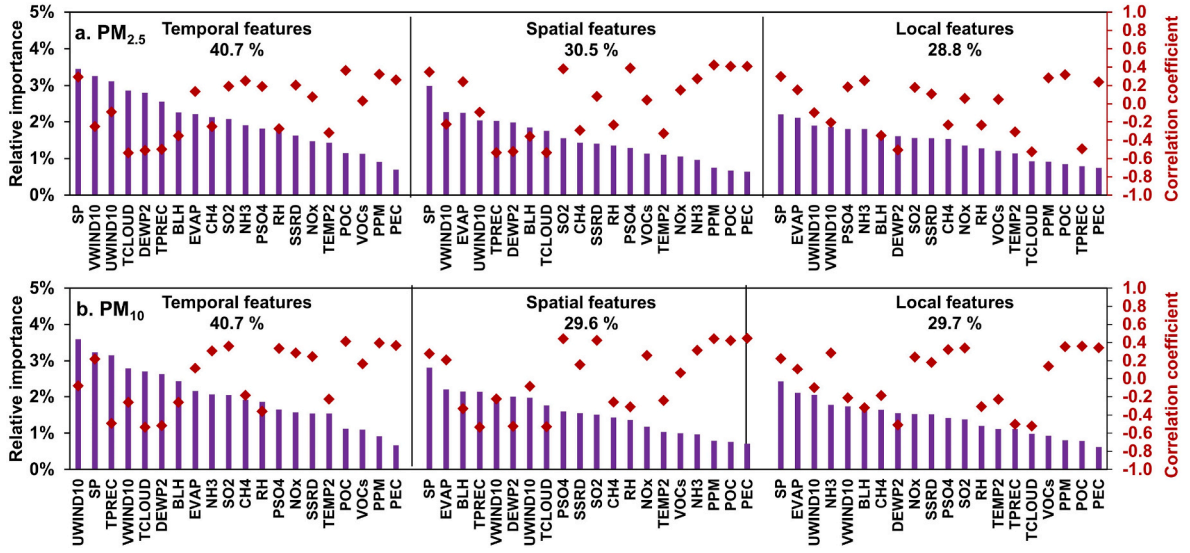


Fig. 2. The features' relative importance and Spearman's correlation coefficient of the $PM_{2.5}$ and PM_{10} estimate models. SSRD: Surface solar radiation, BLH: boundary layer height, EVAP: evaporation, RH: relative humidity, TEMP2: 2 m air temperature, DEWP2: 2 m dewpoint temperature, SP: surface pressure, TPREC: total precipitation, TCLOUD: total cloud cover, UWIND10: 10 m u component of wind, VWIND10: 10 m v component of wind, PPM: primary particulate matter, PSO4: primary sulfate, PEC: primary elemental carbon, POC: primary organic carbon, VOCs: volatile organic compounds, NH_3: Ammonia, NO_X: Nitrogen oxides, SO_2: Sulfur dioxide, CH_4: Methane.

8%, 19%, and 20 % for PM_{10}. The out-of-site CV R^2, RMSE, and MAE (μg/m^3) for the out-of-site CV were 0.65, 28.05 μg m^{-3}, and 46.07 μg m^{-3}, respectively, which are better than the original model. Notably, the extraction of both temporal and spatial features contributes to the model prediction ability. The spatial model shows higher out-of-sample CV accuracy than the original model, with test RMSE and MAE decreased by 1.51 and 1.15 μg/m^3 for $PM_{2.5}$, and by 1.61 and 1.67 μg/m^3 for PM_{10}. The temporal model shows lower RMSE and MAE than the original model, decreased by 7.20 and 4.45 μg/m^3 for $PM_{2.5}$, and by 12.09 and 7.40 μg/m^3 for PM_{10}. Since we cannot access observations in India before 2018, we did not evaluate the performance of the model correspondingly. So instead of using the model to predict unobserved years, we used the model to predict unobserved areas and evaluated the model's spatial predictive performance. The out-of-site CV results show that the R^2 and RMSE of $PM_{2.5}$ are 0.65 and 46.07 μg/m^3, respectively, and there is a 20% drop in model performance compared to the out-of-sample CV, which is mainly due to the out-of-distribution (OOD) problem and can be improved by collecting more observations (Arjovsky, 2020). More importantly, out-of-sample and out-of-site CV of training and testing showed small accuracy gaps reflecting good generalization ability of the spatio-temporal model.

Fig. 3 shows the comparison between ground observations and out-of-sample CV predictions for hourly and daily $PM_{2.5}$ in India. Most data samples of $PM_{2.5}$ and PM_{10} are evenly scattered close to the 1:1 line, with underestimation for high $PM_{2.5}$ and PM_{10} levels. The daily prediction shows better accuracy than hourly prediction, with spatio-temporal R^2, RMSE, and MAE of 0.97, 11.45 μg/m^3, and 8.00 μg/m^3 for $PM_{2.5}$, and 0.98, 19.56 μg/m^3, and 14.19 μg/m^3 for PM_{10}. Hourly validation for $PM_{2.5}$ from 08:00 to 17:00 (Fig. S4) shows high accuracy, with out-of-sample CV R^2 of 0.86–0.88, slope of 0.84–0.86, and intercept of 11.61–12.73 μg/m^3.

Compared to the original model, the spatio-temporal model improves the out-of-sample CV accuracy for hourly and daily $PM_{2.5}$ and PM_{10}. More importantly, the spatio-temporal model has a higher slope (0.84 and 0.95 for hourly and daily $PM_{2.5}$) and lower intercept (12.40 μg/m^3 and 4.23 μg/m^3 for hourly and daily $PM_{2.5}$), indicating improved underestimation of $PM_{2.5}$, especially for high concentration levels.

Several popular regression models (including linear regression, polynomial regression, Random Forest, and XGBoost) were compared with the same features (local and spatial-temporal features), labels, and hyperparameters (Table S3). The tree model performed better than linear and polynomial regression, and XGBoost slightly outperformed lightGBM, but XGBoost has a large gap between training and testing performance (fitted $R^2 = 1$), with the risk of overfitting, whereas the

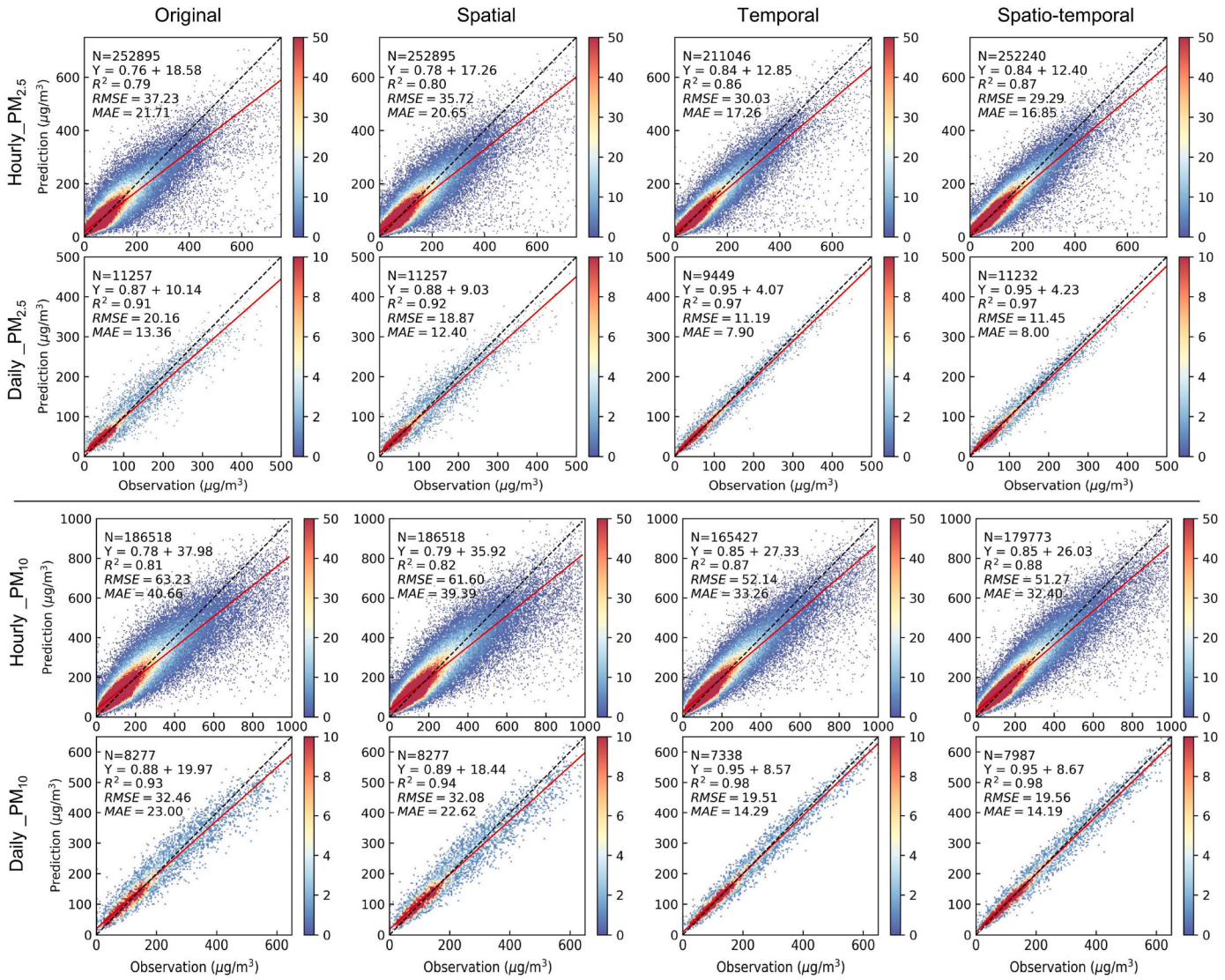


Fig. 3. Comparison between out-of-sample CV prediction and observations for hourly and daily $PM_{2.5}$ and PM_{10}. Dashed lines denote 1:1 line. Solid lines denote linear regression fitting. The sample numbers (N), linear regression function, R^2, RMSE, and MAE are also shown. Units of RMSE and MAE are μg/m^3.

lightGBM model showed smaller gap between training and testing accuracy with faster training speeds.

Since the split ratio between the training and test datasets may also influence the model performance, the model test performance for different test sizes (%) was evaluated (Table S4). The results show that there is no significant difference in model performance when the test set size is below 50%. When the test set is too large, e.g. 90%, there is a loss of model performance due to a small amount of training data.

India has state-based pollution control due to the wide variation in pollution in different parts of India. We have counted the $PM_{2.5}$ pollution by state to understand the $PM_{2.5}$ pollution in different states of India. Fig. 4 shows the $PM_{2.5}$ observations and relative bias of CV predictions compared to observations for the Indian states. Delhi and Uttar Pradesh show severe $PM_{2.5}$ pollution with mean hourly concentrations of 111.2 and 112.0 μg/m^3. Karnataka and Kerala in South India show lower concentration levels (30.8 and 32.9 μg/m^3). Predicted $PM_{2.5}$ concentrations are generally higher than observations. The mean relative bias across Indian states ranges from 17.8% to 66.6%, with median bias ranging from 6.7% to 23.8%. Delhi, Telangana, and Uttar Pradesh show high prediction accuracy, with low median biases and small fluctuations, probably due to the large number and high quality of observations in these regions, which facilitates the ML model to accurately capture the relationship between features and $PM_{2.5}$ concentrations. Karnataka, Punjab, and Rajasthan show large median biases and low accuracy of $PM_{2.5}$ predictions. The Punjab and Rajasthan areas are under the influence of wind-blown dust and show distinct aerosol compositions. Due to limited observations here, the trained model has poor generalization for these areas leading to an overestimation of $PM_{2.5}$ concentrations.

3.3. Predicted regional distribution

The spatio-temporal model was further used to predict maps of $PM_{2.5}$ and PM_{10} (Fig. 5) which are also compared with the CMAQ simulations. Details of the CMAQ simulation description can be found in Supporting. The annual mean concentration across India simulated by CMAQ is 47.4 μg/m^3, and the ML model prediction is 68.3 μg/m^3, which is much closer to the annual mean observation of 68.7 μg/m^3. The spatial distribution of the ML models' predictions is generally consistent with the CMAQ simulations. Due to the limitations of the parameterization for sand generation, there is an underestimation of windblown sand emissions, which further leads to an underestimation in the CMAQ simulation. Most of the observations agree with the predicted values, especially in winter and pre-monsoon season. During the monsoon season, the predicted values are lower than observations in the IGP region of northern India, and CMAQ simulations show the same underestimation. High levels of $PM_{2.5}$ occurred in the Indo-Gangetic Plain (IGP) and western dusty region, with annual mean $PM_{2.5}$ concentrations >90 μg/m^3 in most areas, and lower $PM_{2.5}$ concentrations observed in the southern regions (<40 μg/m^3). The IGP is a fertile alluvial plain with a dense population (>700 million). Intense human activities such as household solid fuel combustion, industries, power plants, vehicles, and construction activities, directly emit large amounts of primary PM and secondary PM precursors (SO_2 and NO_x), and coupled with unfavorable meteorological conditions lead to massive generation and accumulation of $PM_{2.5}$ (Dey et al., 2020). In addition, the Himalayas in the north and the central Indian plateau in the south impede $PM_{2.5}$ dispersion, which instead disperse easterly and westerly under the influence of the monsoon, thus leading to frequent high pollution events in the IGP

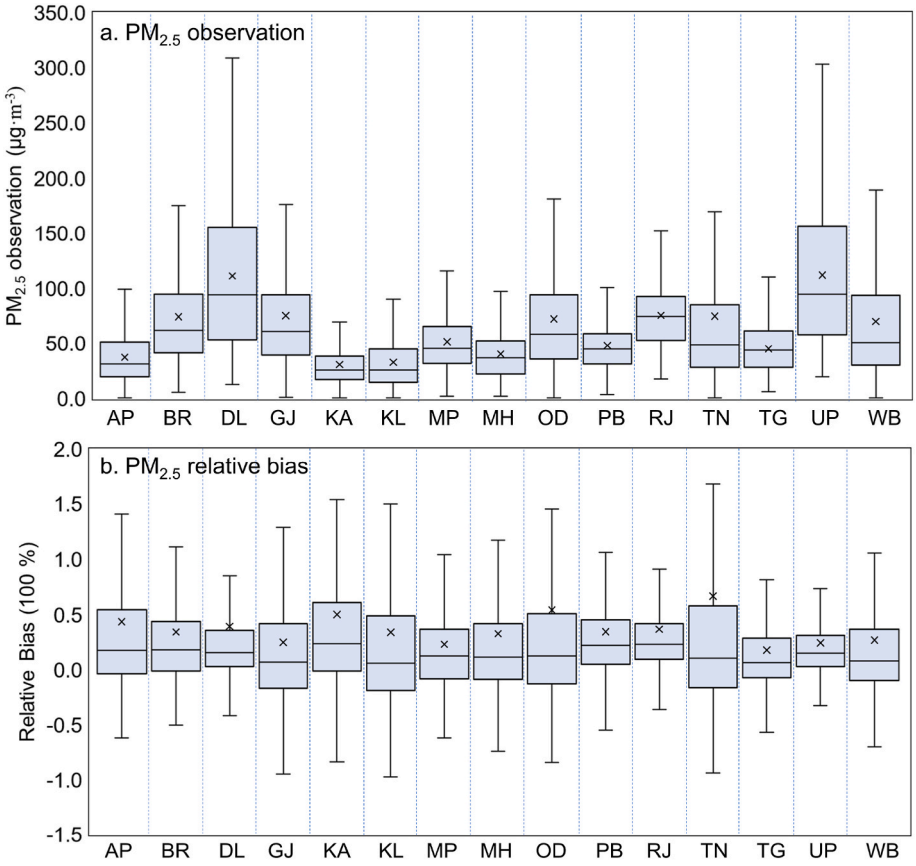


Fig. 4. Boxplots of the hourly $PM_{2.5}$ observations (a) and hourly $PM_{2.5}$ CV prediction relative bias (b) for Indian states and the capital city Delhi. The cross represents the mean bias, and three horizontal lines represent the 25th, 50th, and 75th percentile, respectively. AP: Andhra Pradesh; BR: Bihar; DL: Delhi; GJ: Gujarat; KA: Karnataka; KL: Kerala; MP: Madhya Pradesh; MH: Maharashtra; OD: Odisha; PB: Punjab; RJ: Rajasthan; TN: Tamil Nadu; TG: Telangana; UP: Uttar Pradesh; WB: West Bengal.

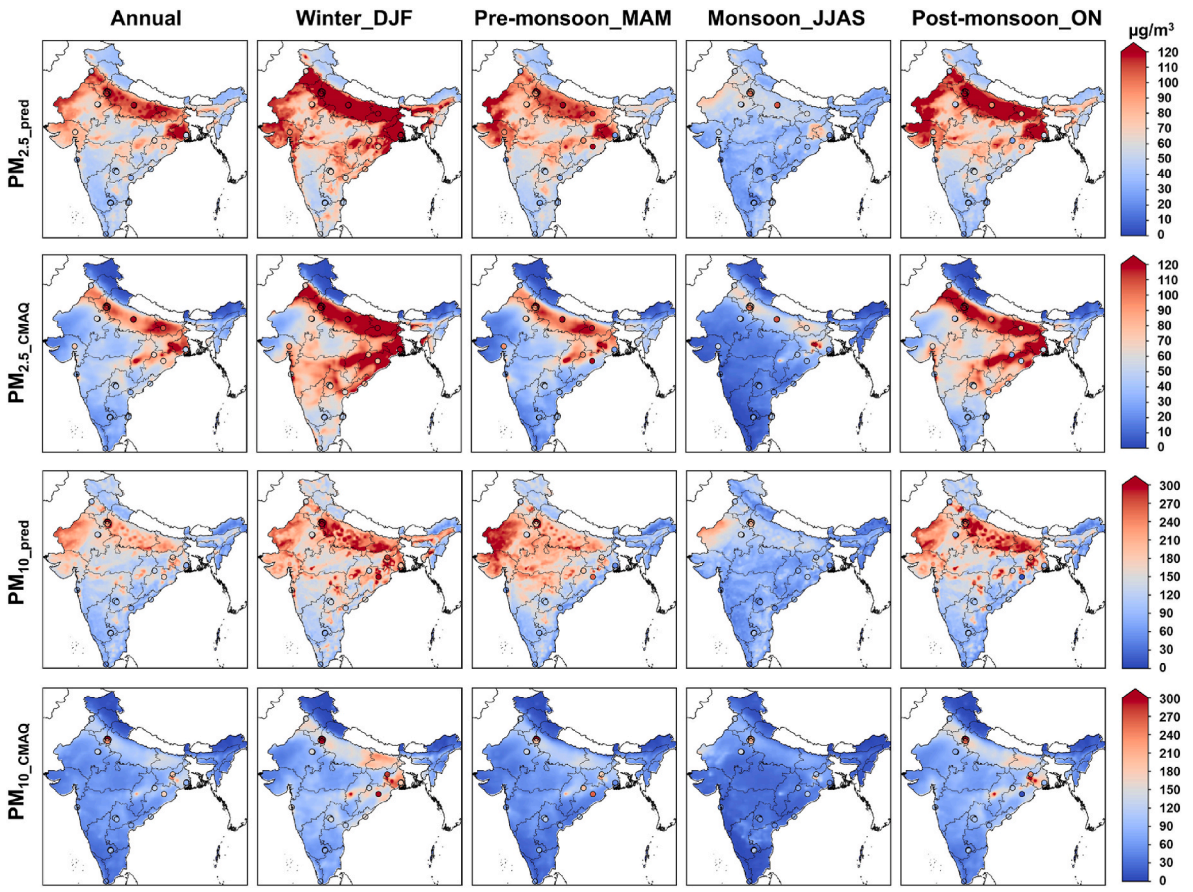


Fig. 5. Annual and seasonal prediction and observations of $PM_{2.5}$ and PM_{10} from ML model and CMAQ simulation in India (2018).

(Maheshwarkar et al., 2022). $PM_{2.5}$ concentrations show a north (high) to south (low) distribution, which is closely related to the distribution of the population and the corresponding anthropogenic emissions.

Seasonally, the most severe $PM_{2.5}$ pollution occurred in winter (96.5 μg/m^3), especially for the IGP region, and the lightest $PM_{2.5}$ pollution occurred in monsoon (41.1 μg/m^3). High $PM_{2.5}$ levels in winter have been attributed to increased emissions from households due to heating and unfavorable stable conditions (Pande et al., 2018). In the monsoon season, the surface wind speed increases with the expansion of the boundary layer height, which facilitates $PM_{2.5}$ dispersion, and reduces $PM_{2.5}$ concentration in highly polluted IGP (Hancock et al., 2023). After the monsoon, the opening biomass burning leads to increased emissions, reduced precipitation leads to weaker scouring, and lower PBL heights lead to weaker dispersion, which combine to increase $PM_{2.5}$ concentrations (Dey et al., 2020). The seasonal trends of $PM_{2.5}$ simulated by CMAQ are consistent with ML predictions, but the spatial distribution shows differences. The PM maps from ML predictions show high concentrations in the dusty region of western India, but the CMAQ simulations fail to reproduce this high PM concentration, which can be attributed to the underestimation of dust aerosols in the CMAQ simulation (Foroutan et al., 2017).

Satellite AOD product MCD19A2 was collected (Fig. S5) for spatial comparison with Monthly prediction of $PM_{2.5}$ (Fig. S6). In the winter months of January November and December, the IGP areas show high $PM_{2.5}$ concentrations which are consistent with high values of AOD. In Odisha, Telangana, and Andhra Pradesh, AOD shows high values from March to June, which is not consistent with low $PM_{2.5}$ concentrations in these regions. This is related to the favorable meteorological and topographical conditions for dispersion (Dey et al., 2020).

3.4. Explanation

Compared to deep learning models, tree models usually have fewer parameters and simpler structures, and the tree relies on only one feature of the Gain per node split, so it is easier to interpret. Meteorological effects on $PM_{2.5}$ are quantified due to its strong influence on $PM_{2.5}$ pollution (Fig. 6) (Chen et al., 2020; Rabha et al., 2021). Emissions were not interpreted due to the large uncertainties in the emission inventories (Crippa et al., 2019; Saikawa et al., 2017). Meteorological normalization was used to separate the meteorological contribution to $PM_{2.5}$ concentrations ($PM_{2.5,meteo}$). SHAP interpretable machine learning method was then used to quantify the contribution of different meteorological factors. Here representative sites from six regions were selected for analysis including, Delhi and Uttar Pradesh (IGP region), Gujarat (Western India region), Maharashtra (Central India region), Odisha (Eastern India region), and Tamil Nadu (Southern India region). Four consecutive days of haze events in winter were screened in order of $PM_{2.5}$ concentration to investigate the meteorological drivers.

Boundary layer height (BLH) is the main meteorological factor influencing $PM_{2.5}$ pollution in several regions. The low BLH in winter is unfavorable for $PM_{2.5}$ dispersion, same as previous studies (Li et al., 2021; Maheshwarkar et al., 2022). The correlation (Table S5) shows a significant negative correlation between BLH and hourly $PM_{2.5}$ concentrations in all regions except the Tamil Nadu region. Tamil Nadu region is located in southern India, and its climate is mainly influenced by the monsoon, and the terrain around the observation site is flattened. The combination of favorable dispersion conditions due to flat topography, strong monsoon influence, and under-representation of observational data resulted in insignificant correlation between BLH and PM2.5 concentrations.

Surface temperature is the main meteorological factor that enhanced

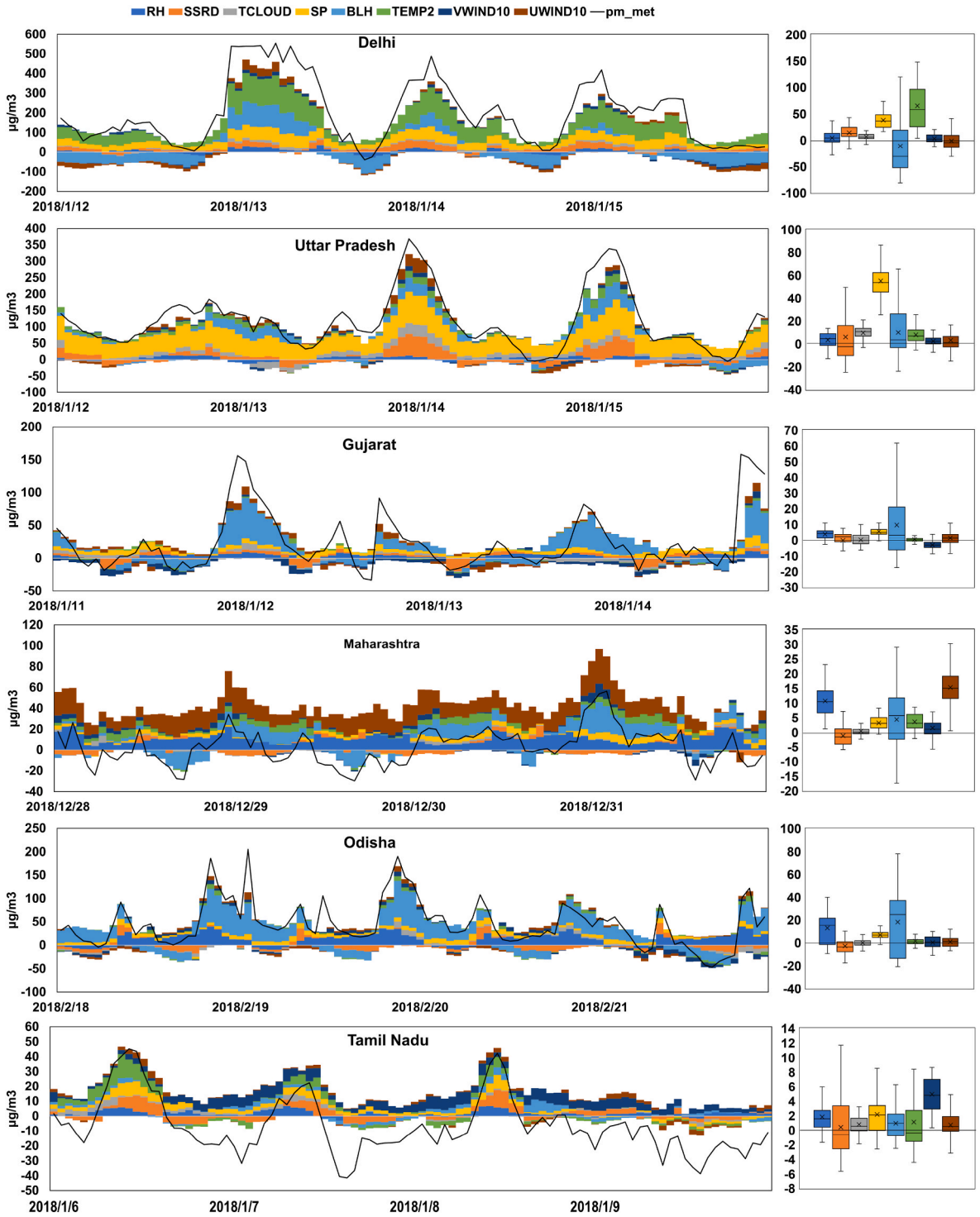


Fig. 6. SHAP values of each meteorological feature for $PM_{2.5}$ and statistic boxplots in six regions of India. Local, spatial, and temporal meteorological features are summed. SSRD: Surface solar radiation, BLH: boundary layer height, RH: relative humidity, TEMP2: 2 m air temperature, SP: surface pressure, TCLOUD: total cloud cover, UWIND10: 10 m u component of wind, VWIND10: 10 m v component of wind, pm_met: meteorological driven $PM_{2.5}$ ($PM_{2.5,meteo}$).

haze formation in Delhi, mainly due to the low temperatures accompanied with high humidity that exacerbated pollution (Fig. S7). In addition, BLH leads to an increase of $PM_{2.5}$ concentration at the onset of the haze event but plays an important attenuating role at the end of pollution, further supporting the important effect of BLH on the dispersion of $PM_{2.5}$. Surface pressure in Uttar Pradesh is the major factor enhancing pollution and is significantly and positively correlated with $PM_{2.5}$ concentration (Table S5). This can be associated with stagnant conditions due to high-pressure atmosphere in winter (Maheshwarkar et al., 2022). The heavy pollution in the Gujarat region is mainly influenced by BLH. The observation site in Gujarat is located in sandy dry area (Fig. S1), which is affected by windblown dust (Chatterjee et al., 2023). The low BLH in winter limits the dispersion of dust aerosols and hence leads to the accumulation of $PM_{2.5}$. Heavy pollution in Maharashtra (Mumbai) was mainly influenced by Relative humidity, U wind, and BLH. U wind shows a significant negative correlation (−0.64) with

$PM_{2.5}$ concentration. In this pollution episode, the mean wind speed in the U direction was −1.57 m/s, implying the effect of regional transport from inland regions on haze pollution, which is in line with the findings of the previous study using back trajectory analysis (Barudgar et al., 2022). Relative humidity enhances haze pollution mainly through increased moisture absorption of particulate matter and accelerated secondary aerosol generation (Maheshwarkar et al., 2022; Rabha et al., 2021). The Tamil Nadu region is mildly polluted and is mainly affected by temperature, radiation, and surface pressure. The Tamil Nadu region is close to the equator with strong surface radiation and the contribution of temperature and radiation may be related to secondary aerosol generation.

3.5. Uncertainties

First, we use a geographic weighting approach to extract regional features, although this method can improve ML model performance, there is still potential to improve it. For example, by combining factors such as topography, population distribution, and atmospheric movement, differential spatial weighting can be achieved to extract more effective regional features using convolutional neural network (Kirkwood et al., 2022). There is also room for improvement in the extraction of temporal features, such as using deep neural networks (Recurrent Neural Networks, RNN; The Long Short-Term Memory, LSTM; and Transformer, etc.) developed in the field of natural language processing to extract more effective features of historical meteorology and emissions (Salehinejad et al., 2017; Vaswani et al., 2017).

Second, the constructed observation dataset is a long-tailed skewed distribution with fewer records of high pollution, so the model tends to underestimate the high pollution scenarios and faces an unbalanced regression problem. This is a common issue reported before and the underestimation was improved by incorporating spatio-temporal features (Wang et al., 2023b; Wei et al., 2021a).

Third, the uncertainty of the input feature sets (ERA5 and EDGAR) also influences the estimation results. ERA5 has been developed over the years and its uncertainties have been systematically analyzed, especially for ERA5-land, which shows good accuracy for most meteorological factors, superior to other reanalyzed data (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). Emission inventories have large uncertainties, due to uncertainties in activity data and emission factors, simplifications in parameterization, regional differences, etc. Emission inventories are one of the key influences in air quality modeling (e.g. CMAQ) (Binkowski and Roselle, 2003; Hu et al., 2016). In tree modeling, uncertainty in the input characteristics can also influence the accuracy of the results, and this influence cannot be measured directly due to the inability to quantify their uncertainty (Ke et al., 2017; Thunis et al., 2024). By the relative importance of features, the impact of emissions uncertainty can be measured indirectly. In this study, the trained model relies heavily on meteorological features (relative contribution of 65%), with emission features (35%) contributing less, and the relative importance of the emission factors is about 2%–6%, which means that 100% of the uncertainty of a single emission feature has 2%–6% impact on the ML model results. By introducing multiple features, the model's reliance on a single feature is reduced, further minimizing the impact of emissions inventory uncertainty on the prediction results.

Finally, the validation of the model was limited to urban observations, as ground-based rural observations were not available, so the accuracy of the generated $PM_{2.5}$ hourly concentration dataset in rural areas is yet to be verified.

Code and data availability

The Python codes of geographic and rolling weighting methods are provided at https://zenodo.org/records/10968469. ERA5 reanalysis data can be reached at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview (last access: 24 August 2023). EDGAR emissions data can be downloaded from https://edgar.jrc.ec.europa.eu/dataset_ap61. CMAQ is an open-source chemical transport model developed by the US Environmental Protection Agency, which can be downloaded at https://github.com/USEPA/CMAQ.

CRediT authorship contribution statement

Shuai Wang: Writing – original draft, Software, Methodology. **Mengyuan Zhang:** Visualization, Validation. **Hui Zhao:** Methodology, Data curation. **Peng Wang:** Writing – review & editing, Methodology. **Sri Harsha Kota:** Data curation. **Qingyan Fu:** Writing – review & editing. **Hongliang Zhang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The Python codes of geographic and rolling weighting methods are provided at https://zenodo.org/records/10968469.

Acknowledgment

This work was supported by the Co-fund DFG-NSFC Sino-German AirChanges project (448720203), the National Natural Science Foundation of China (42077194/42377098), and the Shanghai International Sci-ence and Technology Partnership Project (No. 21230780200).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2024.120834.

References

Anselin, L., Rey, S.J., 2010. Perspectives on spatial data analysis. In: Anselin, L., Rey, S.J. (Eds.), Perspectives on Spatial Data Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–20.

Arjovsky, M., 2020. Out of Distribution Generalization in Machine Learning. New York University.

Barudgar, A., Singh, J., Tyagi, B., 2022. Variability of fine particulate matter (PM2.5) and its association with health and vehicular emissions over an urban tropical coastal station Mumbai, India. Thalassas: Int. J. Mar. Sci. 38, 1067–1080. https://doi.org/10.1007/s41208-022-00442-4.

Binkowski, F.S., Roselle, S.J., 2003. Models-3 community multiscale air quality (CMAQ) model aerosol component 1. Model description. J. Geophys. Res. Atmos. 108. https://doi.org/10.1029/2001JD001409.

Brauer, M., Guttikunda, S.K., K A, N., Dey, S., Tripathi, S.N., Weagle, C., Martin, R.V., 2019. 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.

Chatterjee, D., McDuffie, E.E., Smith, S.J., Bindle, L., van Donkelaar, A., Hammer, M.S., Venkataraman, C., Brauer, M., Martin, R.V., 2023. Source contributions to fine particulate matter and attributable mortality in India and the surrounding region. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.2c07641.

Chen, Z.Y., Chen, D.L., Zhao, C.F., Kwan, M.P., Cai, J., Zhuang, Y., Zhao, B., Wang, X.Y., Chen, B., Yang, J., Li, R.Y., He, B., Gao, B.B., Wang, K.C., Xu, B., 2020. Influence of meteorological conditions on PM2.5 concentrations across China: a review of methodology and mechanism. Environ. Int. 139. https://doi.org/10.1016/j.envint.2020.105558.

Crippa, M., Janssens-Maenhout, G., Guizzardi, D., Van Dingenen, R., Dentener, F., 2019. Contribution and uncertainty of sectorial and regional emissions to regional and global PM2.5 health impacts. Atmos. Chem. Phys. 19, 5165–5186. https://doi.org/10.5194/acp-19-5165-2019.

de Bont, J., Krishna, B., Stafoggia, M., Banerjee, T., Dholakia, H., Garg, A., Ingole, V., Jaganathan, S., Kloog, I., Lane, K., Mall, R.K., Mandal, S., Nori-Sarma, A., Prabhakaran, D., Rajiva, A., Tiwari, A.S., Wei, Y., Wellenius, G.A., Schwartz, J., Prabhakaran, P., Ljungman, P., 2024. Ambient air pollution and daily mortality in

ten cities of India: a causal modelling study. Lancet Planet. Health 8, e433–e440. https://doi.org/10.1016/S2542-5196(24)00114-1.
Dey, S., Purohit, B., Balyan, P., Dixit, K., Bali, K., Kumar, A., Imam, F., Chowdhury, S., Ganguly, D., Gargava, P., Shukla, V.K., 2020. A satellite-based high-resolution (1-km) ambient PM2.5 database for India over two decades (2000–2019): applications for air quality management. Rem. Sens. 12, 3872.
Dhandapani, A., Iqbal, J., Kumar, R.N., 2023. 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.
Foroutan, H., Young, J., Napelenok, S., Ran, L., Appel, K.W., Gilliam, R.C., Pleim, J.E., 2017. Development and evaluation of a physics-based windblown dust emission scheme implemented in the CMAQ modeling system. J. Adv. Model. Earth Syst. 9, 585–608. https://doi.org/10.1002/2016MS000823.
Fotheringham, A.S., Crespo, R., Yao, J., 2015. Geographical and temporal weighted regression (GTWR). Geogr. Anal. 47, 431–452. https://doi.org/10.1111/gean.12071.
Fotheringham, A.S., Oshan, T.M., 2016. Geographically weighted regression and multicollinearity: dispelling the myth. J. Geogr. Syst. 18, 303–329. https://doi.org/10.1007/s10109-016-0239-5.
Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., Huang, X., He, K., Zhang, Q., 2021. Tracking air pollution in China: near real-time PM2.5 retrievals from multisource data fusion. Environ. Sci. Technol. 55, 12106–12115. https://doi.org/10.1021/acs.est.1c01863.
Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? arXiv preprint arXiv:2207.08815.
Guenther, A., Jiang, X., Heald, C.L., Sakulyanontvittaya, T., Duhl, T., Emmons, L., Wang, X., 2012. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2. 1): an extended and updated framework for modeling biogenic emissions. Geosci. Model Dev. (GMD) 5, 1471–1492.
Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P.I., Geron, C., 2006. Estimates of global terrestrial isoprene emissions using MEGAN (model of emissions of gases and aerosols from nature). Atmos. Chem. Phys. 6, 3181–3210.
Guo, H., Kota, S.H., Sahu, S.K., Zhang, H., 2019. Contributions of local and regional sources to PM2.5 and its health effects in north India. Atmos. Environ. 214, 116867. https://doi.org/10.1016/j.atmosenv.2019.116867.
Hammer, M.S., van Donkelaar, A., Li, C., Lyapustin, A., Sayer, A.M., Hsu, N.C., Levy, R. C., Garay, M.J., Kalashnikova, O.V., Kahn, R.A., 2020. Global estimates and long-term trends of fine particulate matter concentrations (1998–2018). Environ. Sci. Technol. 54, 7879–7890.
Hancock, S., Fiore, A.M., Westervelt, D.M., Correa, G., Lamarque, J.-F., Venkataraman, C., Sharma, A., 2023. Changing PM2.5 and related meteorology over India from 1950–2014: a new perspective from a chemistry-climate model ensemble. Environ. Res.: Climate 2, 015003. https://doi.org/10.1088/2752-5295/acb22a.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.
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., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.
Hou, L.L., Dai, Q.L., Song, C.B., Liu, B.W., Guo, F.Z., Dai, T.J., Li, L.X., Liu, B.S., Bi, X.H., Zhang, Y.F., Feng, Y.C., 2022. Revealing drivers of haze pollution by explainable machine learning. Environ. Sci. Technol. Lett. 9, 112–119. https://doi.org/10.1021/acs.estlett.1c00865.
Hu, J., Chen, J., Ying, Q., Zhang, H., 2016. One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. Atmos. Chem. Phys. 16, 10333–10350.
Hu, J., Wang, Y., Ying, Q., Zhang, H., 2014. Spatial and temporal variability of PM2. 5 and PM10 over the north China plain and the yangtze river delta, China. Atmos. Environ. 95, 598–609.
Janzing, D., Minorics, L., Blöbaum, P., 2020. Feature relevance quantification in explainable AI: a causal problem. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2907–2916.
Jin, J., Lin, H.X., Segers, A., Xie, Y., Heemink, A., 2019. Machine learning for observation bias correction with application to dust storm data assimilation. Atmos. Chem. Phys. 19, 10009–10026.
Ke, G.L., Meng, Q., Finley, T., Wang, T.F., Chen, W., Ma, W.D., Ye, Q.W., Liu, T.Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA.
Kirkwood, C., Economou, T., Pugeault, N., Odbert, H., 2022. Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Math. Geosci. 54, 507–531. https://doi.org/10.1007/s11004-021-09988-0.
Li, J., Hao, X., Liao, H., Hu, J., Chen, H., 2021. Meteorological impact on winter PM2.5 pollution in Delhi: present and future projection under a warming climate. Geophys. Res. Lett. 48, e2021GL093722. https://doi.org/10.1029/2021GL093722.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA.
Ma, Z.W., Hu, X.F., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444. https://doi.org/10.1021/es5009399.
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., Jehangir, A., 2022. 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.
Meng, C., Cheng, T.H., Gu, X.F., Shi, S.Y., Wang, W.N., Wu, Y., Bao, F.W., 2019. Contribution of meteorological factors to particulate pollution during winters in Beijing. Sci. Total Environ. 656, 977–985. https://doi.org/10.1016/j.scitotenv.2018.11.365.
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., Thépaut, J.N., 2021. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.
Murray, C.J., Aravkin, A.Y., Zheng, P., Abbafati, C., Abbas, K.M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I.J.T.L., 2020. 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.
Organization, W.H., 2021. WHO Global Air Quality Guidelines: Particulate Matter (PM2. 5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide: Executive Summary.
Oshan, T.M., Smith, J.P., Fotheringham, A.S., 2020. Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. Int. J. Health Geogr. 19. https://doi.org/10.1186/s12942-020-00204-6.
Pande, P., Dey, S., Chowdhury, S., Choudhary, P., Ghosh, S., Srivastava, P., Sengupta, B., 2018. Seasonal transition in PM10 exposure and associated all-cause mortality risks in India. Environ. Sci. Technol. 52, 8756–8763. https://doi.org/10.1021/acs.est.8b00318.
Pant, P., Lal, R.M., Guttikunda, S.K., Russell, A.G., Nagpure, A.S., Ramaswami, A., Peltier, R.E., 2019. Monitoring particulate matter in India: recent trends and future outlook. Air Quality, Atmosphere & Health 12, 45–58.
Ping Tian, D., 2013. A review on image feature extraction and representation techniques. International Journal of Multimedia and Ubiquitous Engineering 8, 385–396.
Qiu, M., Zigler, C., Selin, N.E., 2022. Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. Atmos. Chem. Phys. 22, 10551–10566. https://doi.org/10.5194/acp-22-10551-2022.
Rabha, S., Saikia, B.K., Singh, G.K., Gupta, T., 2021. Meteorological influence and chemical compositions of atmospheric particulate matters in an Indian urban area. ACS Earth Space Chem. 5, 1686–1694. https://doi.org/10.1021/acsearthspacechem.1c00037.
Saikawa, E., Trail, M., Zhong, M., Wu, Q., Young, C.L., Janssens-Maenhout, G., Klimont, Z., Wagner, F., Kurokawa, J.-i., Nagpure, A.S., 2017. Uncertainties in emissions estimates of greenhouse gases and air pollutants in India and their impacts on regional air quality. Environ. Res. Lett. 12, 065002.
Salehinejad, H., Sankar, S., Barfett, J., Colak, E., Valaee, S., 2017. Recent Advances in Recurrent Neural Networks arXiv preprint arXiv:1801.01078.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.
Tec, M., Scott, J.G., Zigler, C.M., 2023. Weather2vec: representation learning for causal inference with non-local confounding in air pollution and climate studies. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14504–14513.
Thunis, P., Kuenen, J., Pisoni, E., Bessagnet, B., Banja, M., Gawuc, L., Szymankiewicz, K., Guizardi, D., Crippa, M., Lopez-Aparicio, S., Guevara, M., De Meij, A., Schindlbacher, S., Clappier, A., 2024. Emission ensemble approach to improve the development of multi-scale emission inventories. Geosci. Model Dev. (GMD) 17, 3631–3643. https://doi.org/10.5194/gmd-17-3631-2024.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, K.C., Dickinson, R.E., Wild, M., Liang, S., 2012. Atmospheric impacts on climatic variability of surface incident solar radiation. Atmos. Chem. Phys. 12, 9581–9592. https://doi.org/10.5194/acp-12-9581-2012.
Wang, S., Kota, S.H., Zhang, H., 2023a. LongPMInd: Long-Term (1980-2022) Daily Ground Particulate Matter Datasets in India. Zenodo.
Wang, S., Wang, P., Qi, Q., Wang, S., Meng, X., Kan, H., Zhu, S., Zhang, H., 2023b. Improved estimation of particulate matter in China based on multisource data fusion. Sci. Total Environ., 161552
Wang, S., Wang, P., Zhang, R., Meng, X., Kan, H., Zhang, H., 2023c. Estimating particulate matter concentrations and meteorological contributions in China during 2000–2020. Chemosphere 330, 138742. https://doi.org/10.1016/j.chemosphere.2023.138742.
Wang, S., Zhang, M., Zhao, H., Wang, P., Kota, S.H., Fu, Q., Liu, C., Zhang, H., 2024. Reconstructing long-term (1980–2022) daily ground particulate matter concentrations in India (LongPMInd). Earth Syst. Sci. Data 16, 3565–3577. https://doi.org/10.5194/essd-16-3565-2024.
Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021a. Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Rem. Sens. Environ. 252, 112136.

Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., Zhu, T., 2023. First close insight into global daily gapless 1 km PM2.5 pollution, variability, and health impact. Nat. Commun. 14, 8349. https://doi.org/10.1038/s41467-023-43862-3.

Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W., Li, R., Cribb, M., 2021b. Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). Atmos. Chem. Phys. 21, 7863–7880. https://doi.org/10.5194/acp-21-7863-2021.

Wiedinmyer, C., Akagi, S.K., Yokelson, R.J., Emmons, L.K., Al-Saadi, J.A., Orlando, J.J., Soja, A.J., 2011. The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning. Geosci. Model Dev. (GMD) 4, 625–641. https://doi.org/10.5194/gmd-4-625-2011.

Xiao, Q., Zheng, Y., Geng, G., Chen, C., Huang, X., Che, H., Zhang, X., He, K., Zhang, Q., 2021. Separating emission and meteorological contributions to long-term PM2.5 trends over eastern China during 2000–2018. Atmos. Chem. Phys. 21, 9475–9496. https://doi.org/10.5194/acp-21-9475-2021.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., Wang, W., 2016. National-scale estimates of ground-level PM2. 5 concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. Rem. Sens. 8, 184.

Zhao, W.Z., Du, S.H., 2016. Spectral-spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. IEEE Trans. Geosci. Rem. Sens. 54, 4544–4554. https://doi.org/10.1109/tgrs.2016.2543748.