MACHINE LEARNING
Science and Technology

PURPOSE-LED
PUBLISHING™

PAPER • OPEN ACCESS

DeepAir: deep learning and satellite imagery to estimate high-resolution $PM_{2.5}$ at scale

To cite this article: Wenxuan Guo *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 015057

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
5 August 2024

REVISED
26 January 2025

ACCEPTED FOR PUBLICATION
14 February 2025

PUBLISHED
4 March 2025

PAPER

DeepAir: deep learning and satellite imagery to estimate high-resolution $PM_{2.5}$ at scale

Wenxuan Guo[1,5], Zhaoping Hu[1,5], Ling Jin[2], Yanyan Xu[1,2,3,*] and Marta C Gonzalez[2,3,4,*]

[1] MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
[2] Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States of America
[3] Department of City and Regional Planning, University of California, Berkeley, CA 94720, United States of America
[4] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, United States of America
[5] These authors contributed equally.
[*] Authors to whom any correspondence should be addressed.

E-mail: yanyanxu@sjtu.edu.cn and martag@berkeley.edu

Keywords: high resolution $PM_{2.5}$, spatial interpolation, satellite imagery, deep learning

Abstract

Air pollution, specifically $PM_{2.5}$, has become a significant global concern owing to its detrimental impacts on public health. Even so, the high-resolution monitoring of air pollution is still a challenge on a global scale. To cope with this, machine learning (ML) techniques have been utilized to infer the concentration of air pollutants at a fine scale. In this study, we propose *DeepAir*, a learning framework for estimating $PM_{2.5}$ concentrations at a fine scale with sparsely distributed observations. *DeepAir* integrates a pre-trained convolutional neural network with the LightGBM method. This framework estimates the $PM_{2.5}$ concentration of a given patch, utilizing a synergy of geographical information, meteorological conditions, and satellite observations. We select California as the focal region and train the model with data from 2014 to 2017 provided by 130 $PM_{2.5}$ observation stations in the state. Upon training, the model can be applied to estimate the daily $PM_{2.5}$ concentrations at 1 km resolution across California. Our methodology meticulously incorporates meteorological variables, with a particular emphasis on wildfire propagation, and contemplates the complex interplay of various features. To ascertain the efficacy of our model, we employ the 10-fold cross-validation technique, which confirms that our model surpasses traditional ML and standalone deep learning methods.

1. Introduction

Air pollution is a global issue that severely affects human health and the ecological environment [1]. With the acceleration of industrialization and urbanization, it has become an increasingly prominent problem in many countries and regions. Various pollutants such as particulate matter (PM), nitrogen oxides, sulfur oxides, and heavy metal ions directly impact vulnerable populations. Prolonged exposure to heavily polluted environments can lead to various respiratory diseases, cardiovascular diseases, and even malignancies of grave severity. Among these pollutants, fine PM ($PM_{2.5}$)—particles less than 2.5 μm in diameter—poses a particular threat. Due to their small size, $PM_{2.5}$ particles can be inhaled deeply into the respiratory tract, irritate and corrode the alveolar wall, and consequently impair lung function [2]. Chronic exposure to $PM_{2.5}$ is linked to severe health conditions such as asthma, chronic obstructive pulmonary disease, cardiovascular issues, and increased mortality rates. $PM_{2.5}$ tend to accumulate more than PM_{10}, propagate long distances, become stagnant in the atmosphere, stay in the air longer, and travel farther [3, 4]. According to [5], a 10 $\mu g\,m^{-3}$ increment in $PM_{2.5}$ was associated with a 1.04% increase in the risk of death. Therefore, studying $PM_{2.5}$ is crucial for understanding and mitigating its specific health impacts on urban populations. To combat this problem, governments worldwide are establishing atmospheric pollution monitoring stations to enable real-time detection of pollutant concentrations, thereby supporting the development of effective environmental management strategies [6, 7]. Additionally, by making air quality data accessible to the public, residents gain a clearer understanding of air quality levels and receive timely warnings about pollution events. This empowers individuals to make informed lifestyle choices, such as adjusting outdoor activities or adopting protective measures like wearing masks.

However, air pollution monitors are often sparsely distributed due to the high equipment costs and land use policy restrictions. This limited coverage necessitates reliable prediction models to estimate pollutant concentrations across unmonitored areas. To enhance prediction accuracy, researchers have investigated the causal relationships between air pollution and various environmental factors, including both natural and built environments [8]. In particular, meteorological data play a significant role in shaping air quality. Variables such as temperature, humidity, wind speed, and wind direction exert substantial influence on the dispersion and accumulation of pollutants [9]. Under certain meteorological conditions, pollution levels can intensify, while in other conditions, these variables aid in dispersing pollutants, reducing their concentration in the air. In addition to meteorological data, satellite-based observations, such as aerosol optical depth (AOD), have become invaluable in assessing air quality and quantifying pollution levels [10–12]. AOD measures the concentration of aerosols in the atmosphere, reflecting the cumulative aerosol content within a vertical column, which is a critical metric in pollution studies.

Recognizing the strong correlations among air pollution, meteorological factors, and satellite observations, researchers have increasingly sought to spatially infer pollutant concentrations by integrating both meteorological and satellite data. By combining these dynamic data sources with static geographic information, such as elevation and land use types, it becomes possible to estimate pollutant levels across a broader area, even in locations where direct monitoring is absent. In recent years, machine learning (ML) techniques have shown great promise in estimating air pollutant concentrations in regions lacking direct monitoring [13]. Due to their flexibility and adaptability, ML methods are well-suited to this task and have been widely explored in previous research.

Early studies often rely on traditional statistical approaches, including the land use regression (LUR) model, which predicts pollutant concentrations by modeling linear relationships between environmental characteristics and pollution levels [14–16]. Other works employ generalized additive models [11, 17, 18] to model the relationships among air pollution, meteorological factors, satellite observations, and $PM_{2.5}$ pollutant concentrations. Non-parametric regression models such as Random forest [19], XGBoost [20], and LightGBM [21] have also been applied, leveraging meteorological and satellite data to infer pollutant levels. For instance, Hu *et al* [22] used a Random forest model to estimate $PM_{2.5}$ concentrations in the United States. Additionally, mixed effects models [12, 23, 24] have been employed to incorporate the heterogeneity in environmental data, enhancing predictive accuracy by incorporating both meteorological and satellite observations.

With advancements in deep learning, neural network-based models have been developed to predict spatial $PM_{2.5}$ concentrations [25–27]. We pay special attention to works utilizing CNNs as backbones [28–30], which is a key component in our framework. They use more complex modeling approaches for modeling real photographic data [28] or city-level spatial dependencies [30] using spatio-temporal attention convolutional neural networks. [29] addresses pollutant concentration prediction with spatio-temporal attention convolutional neural networks.

Traditional ML models like LUR, Random forest, and LightGBM offer efficient inference and relatively simple model structures, making them advantageous in terms of computational speed. However, they often require manual feature engineering and are limited in capturing complex patterns in high-dimensional data. Furthermore, they solely leverage the features within the target cell when predicting pollutant concentrations, ignoring the critical role of spatial adjacency in this task. Deep learning models, on the other hand, can automatically learn intricate relationships among meteorological data, satellite observations, and pollutant concentrations, making them more accurate in capturing spatio-temporal dependencies. Nevertheless, deep learning models are computationally intensive, requiring extensive training data and facing scalability challenges due to slower inference speeds. Existing works [28–30] typically focus on predicting hourly pollutant concentrations at a limited number of monitoring stations, which fundamentally differs from our objective of estimating daily average $PM_{2.5}$ levels across California at a fine spatial resolution. Given this scope, computational efficiency is paramount, and the complexity of existing CNN-based methods limits their applicability to large-scale, high-resolution tasks.

To address these limitations and combine the advantages of both approaches, we propose a hybrid model named *DeepAir*. As illustrated in figure 1, DeepAir integrates a pre-trained convolutional neural network for feature encoding with LightGBM to capture valuable spatio-temporal information for fast $PM_{2.5}$ concentration inference. Compared to existing models, DeepAir offers significant advantages, including finer spatiotemporal resolution, coverage of a larger study area, and the ability to achieve large-scale inference using observations from limited monitoring stations. Additionally, the model achieves a notable speedup through its use of a pretrained architecture and incorporates novel feature engineering, such as integrating

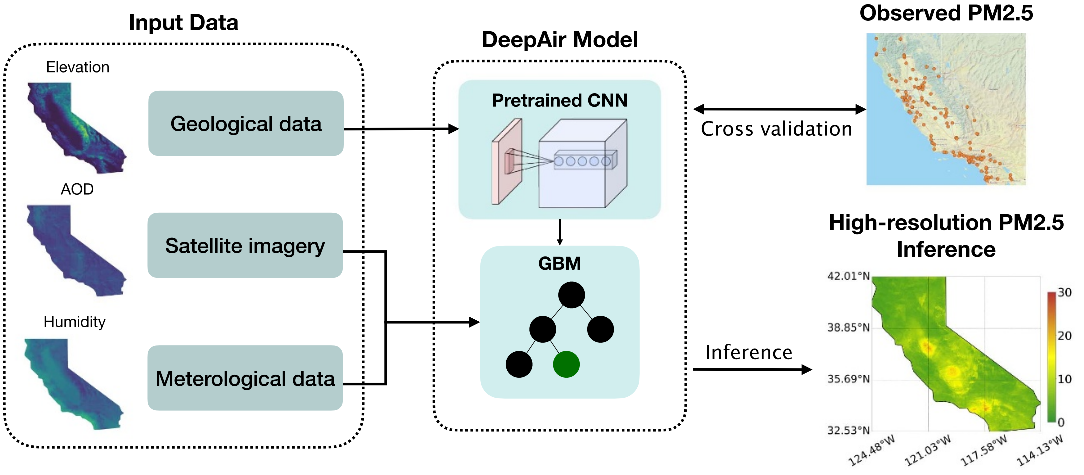

Figure 1. The pipeline of the DeepAir model. DeepAir's input variables include satellite observations (e.g. AOD), static geographic information (e.g. elevation), and meteorological information (e.g. humidity). Trained on $PM_{2.5}$ data from air quality monitors across California, DeepAir is capable of inferring $PM_{2.5}$ concentrations across the entire state at a fine spatial resolution of 1 km.

data from neighboring monitoring stations and leveraging a physical model to simulate wildfire impacts. These innovations enable DeepAir to provide an efficient and scalable solution for high-resolution $PM_{2.5}$ estimation, setting it apart from existing approaches. Extensive experiments validate the effectiveness and efficiency of our approach, with 10-fold cross-validation on real-world data from California. Results show that DeepAir outperforms both traditional ML models and standalone deep learning models, demonstrating superior predictive accuracy and inference efficiency.

2. Methodology

2.1. Preliminary

2.1.1. Study domain

This study centers on the state of California in the United States, utilizing daily observed $PM_{2.5}$ concentrations collected from 130 air quality monitoring stations across California between 2014 and 2017. In addition, statewide meteorological data and satellite observation data are incorporated into the analysis. The spatial distribution of air quality monitoring stations in California is depicted in figure 2(a). Figure 2(b) illustrates the temporal variations in selected meteorological variables (including evaporation, humidity, and wind speed in specific directions) and satellite observation AOD alongside $PM_{2.5}$ concentration recorded at a monitoring station in the San Francisco Bay Area in 2016. Notably, around October 2016, wildfires in the Bay Area led to anomalous spikes in these observed data. These fluctuations highlight the correlations between $PM_{2.5}$ levels, meteorological conditions, and satellite observation data.

2.1.2. Data

The DeepAir model estimates $PM_{2.5}$ concentrations by integrating meteorological data, satellite observations, and static environmental data across the entire California region. These data are recorded for each 1 km × 1 km spatial cells, with a total of 438 619 cells that cover California from 2014 to 2017. For model training, we utilize daily $PM_{2.5}$ concentration data from 130 air quality monitoring stations throughout California. These daily observations are provided by the United States Environmental Protection Agency [6]. Figure 3 visualizes the average value in 2016 of some meteorological data and satellite observation data used for prediction. In particular, the 'Emission' data in the raw dataset is recorded annually, whereas the other variables are recorded daily. The remainder of this section offers a detailed description of the static environmental data, meteorological data, and satellite observation data utilized in the DeepAir model.

Static data. This study leverages data from 2014 to 2017. In this time span, the elevation and land use type (land cover) is assumed to remain constant within each cell in California, constituting static geographical information fed to the pre-trained CNN module in our model. Additionally, each 1 km × 1 km cell contains data on the distance to roads and annual average emission data, which are integrated as features for inferring $PM_{2.5}$ concentrations in subsequent steps.

[6] https://www.epa.gov/outdoor-air-quality-data.

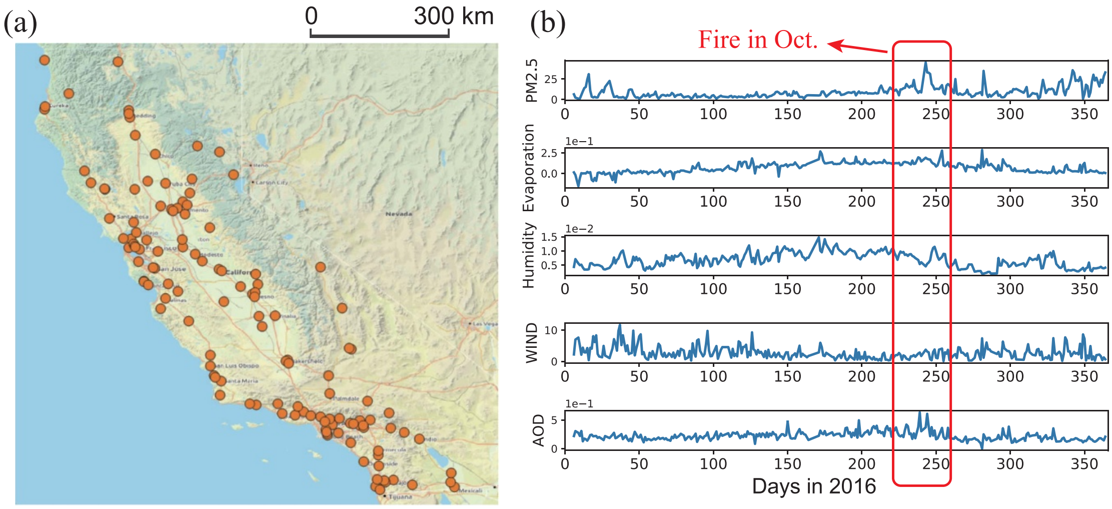

Figure 2. (a) Locations of $PM_{2.5}$ monitoring stations in California. (b) Yearly variations in $PM_{2.5}$ concentrations, meteorological variables, and satellite observations in the San Francisco Bay Area, California, during 2016.

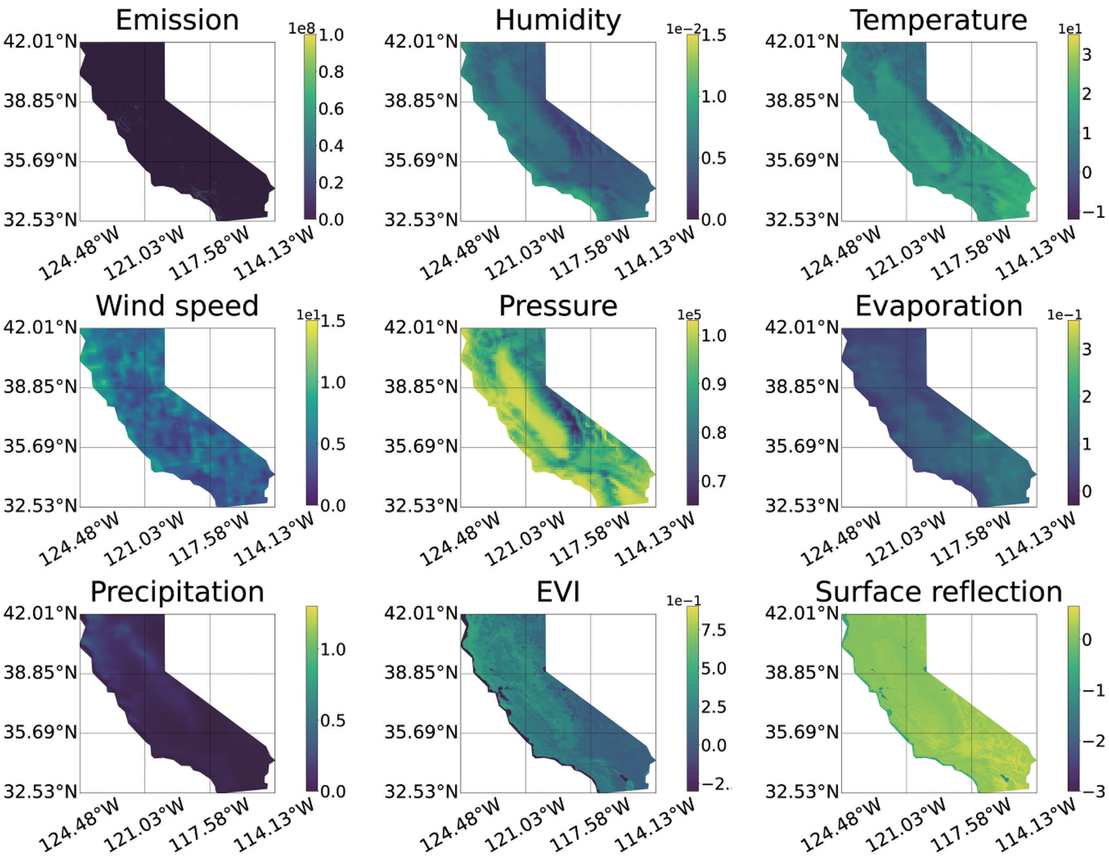

Figure 3. Example of meteorological data and satellite observation data in California. Emissions are recorded annually and others are daily averages in 2016.

Meteorological data. As depicted in figure 2(b), meteorological conditions have a marked correlation with $PM_{2.5}$ levels. Variables included in the DeepAir model encompass 'Temperature', 'Humidity', 'Pressure', 'uWind', 'vWind', 'Evaporation', 'Precipitation', and wildfire data, where 'uWind' and 'vWind' represent eastward and northward components of wind vectors. These daily averages, spanning 2014–2017, are sourced from the North American Land Data Assimilation System [7] developed by NASA, with a spatial

[7] https://ldas.gsfc.nasa.gov/nldas.

resolution of 0.125° [31]. During preprocessing, selected meteorological features were refined further, as elaborated in subsequent sections.

Satellite observation data. For satellite observations, we mainly leverage 'EVI' (enhanced vegetation index) data, 'SR' (surface reflection) data and 'AOD' data for California. The data were collected and initially processed by NASA [8]. Specifically, EVI data are extracted from the MODIS measurements (MOD13A2) [9], provided every 16 days at a 1 km spatial resolution and used to monitor vegetation and land cover changes. SR data, sourced from MODIS (MOD09A1) [10], provides an estimate of the surface spectral reflectance every 8 days at a 0.5 km resolution. AOD data are extracted from MCD19A2 (MODIS/Terra+Aqua Land AOD Daily L2G Global 1 km SIN Grid) [11]. The MCD19A2 product is a Moderate Resolution Imaging Spectroradiometer (MODIS) Terra and Aqua combined Multi-angle Implementation of Atmospheric Correction (MAIAC) Land AOD gridded Level 2 product produced daily at 1 km pixel resolution. Similar to meteorological data, preprocessed satellite data are converted to daily averages for each 1 km × 1 km cell across California, covering 2014–2017.

Land cover data. Land cover significantly influences $PM_{2.5}$ concentration, as different land types affect the sources, dispersion, and removal processes of PM [17]. For example, urban areas, characterized by dense populations and industrial activity, typically have higher $PM_{2.5}$ emissions from vehicles, factories, and construction sites. In contrast, forests can act as both sources and sinks for $PM_{2.5}$—emitting particles through natural processes while also reducing concentrations via vegetation capturing and deposition on leaf surfaces. Land cover data is derived from the National Land Cover Database 2016, which provides fine-scale (30 m) land use information, developed by the U.S. Geological Survey (USGS) [12].

Elevation data. Elevation is another factor influencing $PM_{2.5}$ dispersion. Higher elevations typically exhibit lower air pressure and density, affecting the dispersion and dilution of $PM_{2.5}$ particles. We utilize high-resolution elevation data provided by the U.S. Geological Survey to enhance model accuracy.

CO_2 emission data. Vehicle emissions also significantly contribute to $PM_{2.5}$ concentrations, and on-road CO_2 emissions are therefore considered in our model [32]. The Oak Ridge National Laboratory Distributed Active Archive Center openly provides annual CO_2 emissions data at 1 km resolution for the contiguous United States [33]. These emissions are calculated based on roadway-level vehicle traffic data and state-specific emission factors for various vehicle types on both urban and rural roads.

2.1.3. Data preprocessing

In order to better infer $PM_{2.5}$ concentration spatially, we preprocessed the original meteorological data and satellite observation data combined with $PM_{2.5}$ concentration data provided by air pollution monitoring stations.

AOD imputation. The original satellite observation data primarily comprises AOD values recorded daily from 2014 to 2017. However, this dataset contains spatial gaps in the coverage of AOD. Missing AOD data can significantly hinder the model's ability to accurately predict ground-level $PM_{2.5}$ concentrations and may introduce unavoidable biases [34]. In response, previous studies [35] have commonly employed imputation techniques to fill these gaps before proceeding with $PM_{2.5}$ predictions. Following this approach, a key component of our feature engineering process is the imputation of missing AOD values in the original dataset.

As previously noted, AOD has a high correlation with certain meteorological factors and pollutant concentrations. Given that our meteorological data is complete and does not contain any missing values across both temporal and spatial dimensions, we leveraged LightGBM to develop a regression model specifically for AOD imputation. This model predicts AOD for specific locations by utilizing meteorological variables, including temperature, humidity, pressure, evaporation, and precipitation, as well as static geographical information, such as elevation.

The dataset used for training includes about 16 000 spatial cells with AOD records from 2014 to 2017. We used a 70–30 train-test split to validate the model's performance. The model achieved an imputation R^2 exceeding 94%, demonstrating strong predictive reliability. Subsequently, we applied this trained model to

[8] https://data.nasa.gov/.

[9] http://doi.org/10.5067/MODIS/mod13a2.006.

[10] https://lpdaac.usgs.gov/products/mod09a1v006/.

[11] https://lpdaac.usgs.gov/products/mcd19a2v006/.

[12] https://www.mrlc.gov/data/nlcd-2016-land-cover-conus.

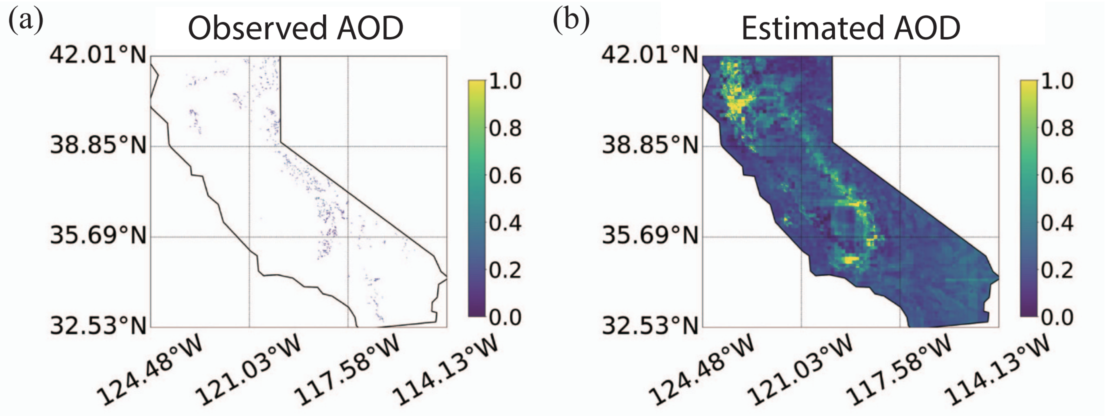

Figure 4. AOD imputation. (a) Original AOD data on 1, July 2016. (b) AOD data after imputation.

complete the missing AOD data in the satellite observation dataset. Figure 4 provides a visualization of both the original and imputed AOD values across California on 1 July, 2016.

Wildfire modeling. Wildfires are significant sources of $PM_{2.5}$, not only increasing pollution levels within affected cells but also impacting surrounding regions due to the dispersion of smoke and particles by wind. The original dataset records wildfire occurrences on a daily basis, using a binary indicator (0 or 1) for each 1 km × 1 km cell. However, this data only documents the initial location and timing of wildfires, without considering their impact on surrounding cells. Since $PM_{2.5}$ data originates from cells containing air quality monitoring stations, where wildfires are relatively rare, directly using this binary wildfire data may limit the model's ability to capture the correlation between wildfire events and $PM_{2.5}$ concentrations. To accurately model this diffusion effect, we take wind speed and direction into account. The original meteorological dataset includes 'uWIND' and 'vWIND' variables, representing wind vectors in the zonal and meridional directions, respectively. We use these vectorized data to calculate the original wind speed and direction at each cell, which then serve as inputs for simulating wildfire spread.

To model wildfire propagation, we treat the cell where the wildfire originates as the starting point of a piriform (teardrop-shaped) curve, representing the theoretical area impacted by the fire. This curve is defined by:

$$x = -b \times (\sin\theta - 1) \tag{1}$$

$$y = a \times (1 - \sin\theta) \times \cos\theta \tag{2}$$

where parameters a and b control the shape of the curve, and θ determines its orientation. Figure 5(a) illustrates the simulated spread of a wildfire occurring in the grid at (122.13° W, 37.67° N).

For cells within the area enclosed by the piriform curve, we further calculate a wildfire impact index based on restored wind speed v_{wind} and direction data. The wildfire propagation is simulated utilizing the actual wind direction as the axis, with the wildfire impact index w formulated as follows:

$$\text{range} = v_{\text{wind}} \times 4/100, \tag{3}$$

$$a = 0.6 \times \text{range}, \tag{4}$$

$$b = \text{range}, \tag{5}$$

$$v_\theta = v_{\text{wind}} \times \cos^2\theta, \tag{6}$$

$$w = \frac{\sqrt{v_\theta / v_{\text{base}}}}{e^{\text{dist}}} \times \frac{1 - \text{dist}}{\text{range} + \epsilon} \times \cos^2\theta. \tag{7}$$

In these equations, range is a wind-speed-dependent parameter that determines the extent of the wildfire-affected area. The angle θ denotes the angle between each cell in the affected region and the wildfire origin cell, relative to the wind axis. The term v_θ represents the wind speed component in this direction, modulating the wildfire's spread capability, and *dist* is the distance between each cell and the cell where the fire originated (in degree). Finally, the wildfire impact index, w, is calculated for each cell within the affected area. During preprocessing, only cells with $w > 10^{-4}$ are retained. The parameter v_{base} is set to 40 to scale the wildfire's influence on adjacent regions. Figure 5(b) displays the computed impact weights for neighboring cells based on the modeled spread.

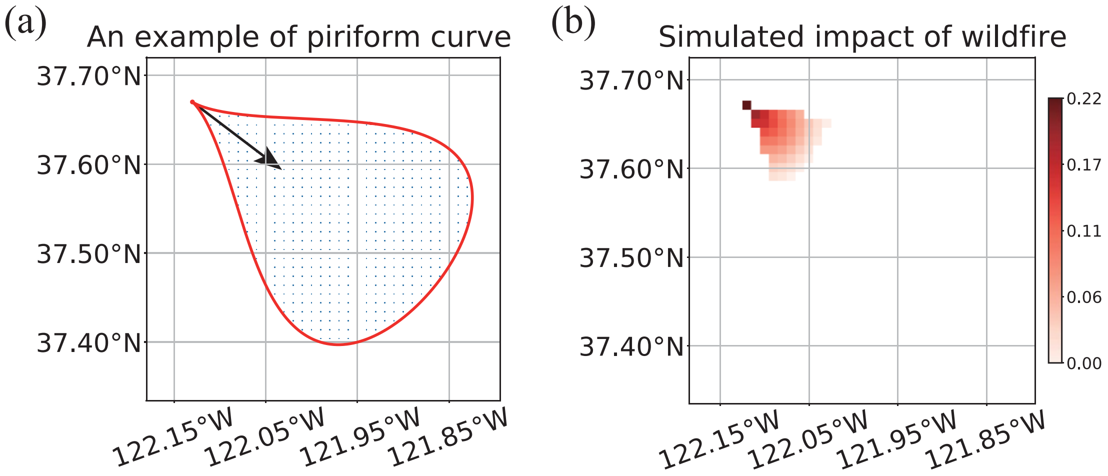

Figure 5. Wildfire spread simulation. (a) Simulation of wildfire spread areas. (b) Distribution of wildfire impact weights.

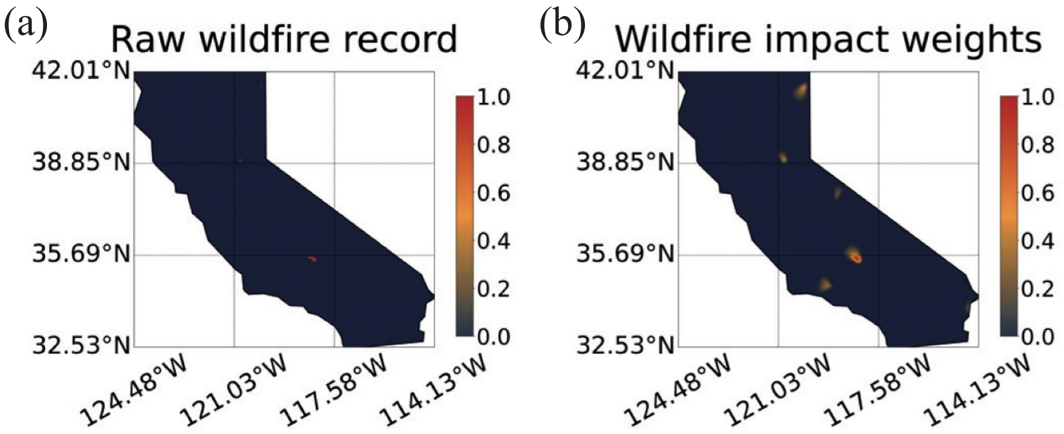

Figure 6. Results of wildfire simulation on 1 July, 2016 in California. (a) Original records of wildfire occurrences in California. (b) Distribution of wildfire impact weights derived from simulated wildfire spread.

Following this approach, we processed all wildfire records in the dataset. figure 6(a) presents the raw wildfire data as binary indicators for California on July 1, 2016. Figure 6(b) shows the computed wildfire impact weights after diffusion modeling, which effectively captures the influence of wildfires on surrounding areas.

New feature from the nearby air quality monitoring stations. After processing the original AOD and wildfire data, we enhance the dataset by creating new features from neighboring air quality monitoring stations. Given the spatial coherence of $PM_{2.5}$ concentrations, $PM_{2.5}$ data from nearby cells provide auxiliary insights for modeling the concentration in a target cell. Therefore, we construct new features using recorded data from nearby monitoring stations for better inference. Specifically, we identify the five nearest monitoring stations for each cell, recording their $PM_{2.5}$ levels $\{\text{nb\_PM2.5}_i\}_{i=1,\ldots,5}$ and distances $\{\text{nb\_Dist}_i\}_{i=1,\ldots,5}$ from the target cell. We then construct features $[\text{nb\_PM2.5}_i/\text{nb\_Dist}_i]_{i=1,\ldots,5}$ to quantify the influence of these nearby monitoring stations.

Additional data preprocessing. We have so far acquired imputed AOD data, simulated wildfire impact weights and nearby air quality data for each cell following the aforementioned feature processing. Additional feature transformations and constructions on the original environmental monitoring data are as follows:

- Deleting outliers from the $PM_{2.5}$ concentration records, specifically, records with daily average $PM_{2.5}$ concentrations less than 0.5 $\mu g\,m^{-3}$ or greater than 400 $\mu g\,m^{-3}$.
- Assigning the corresponding season based on the month of the $PM_{2.5}$ records.
- Calculating the actual wind speed and direction based on the values of ‘uWIND’ and ‘vWIND’.
- Computing the wildfire impact weights for the five nearest air quality monitoring stations of every cell.

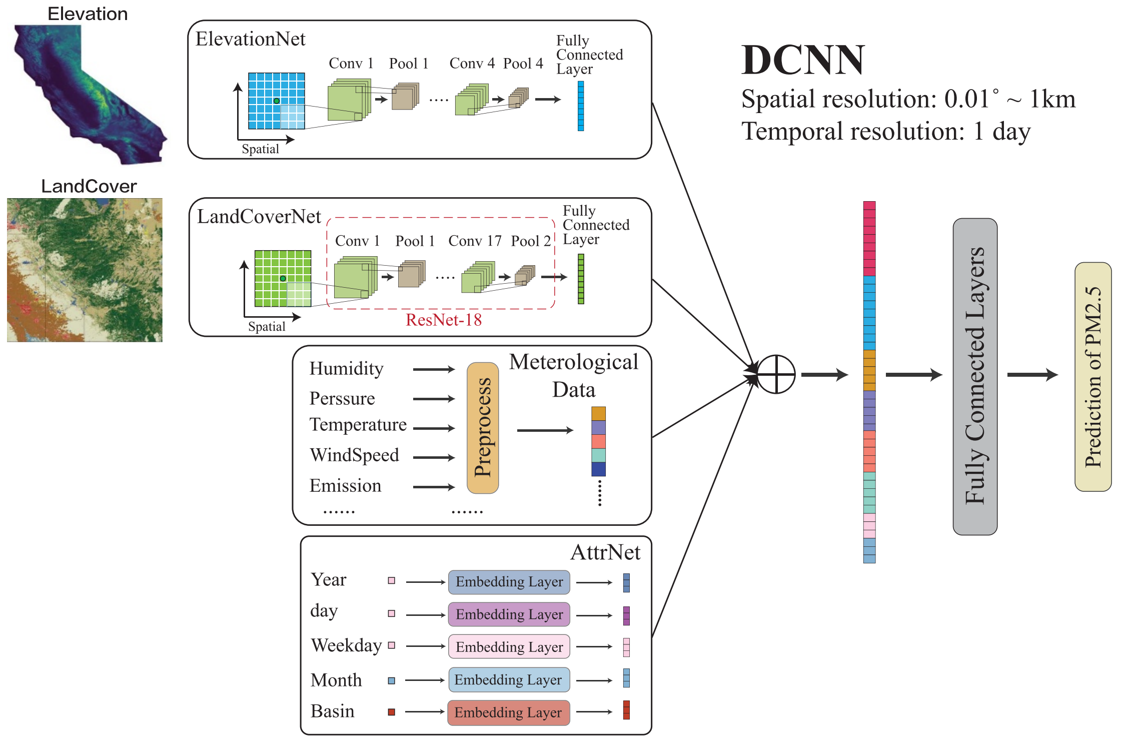

Figure 7. Architecture of the DCNN model.

2.2. Model architecture

Our proposed model, DeepAir, combines a pre-trained deep convolutional neural network (DCNN) and gradient boosting machine (GBM) for effective and efficient inference of $PM_{2.5}$ concentrations at scale. Compared to a simple concatenation of features of the target cell, CNN has the advantage of aggregating information from spatially adjacent areas. Considering the computation overhead of CNN, we use static geographical (elevation and land use type data) as its inputs. We first introduce the CNN modules for feature encoding, followed by two implementations, namely DCNN and DeepAir.

CNN Module for Feature Extraction Elevation and land use type are two important geographical factors that can affect the concentration of $PM_{2.5}$. In this paper, both features are assumed constant during the studied time span. We design two sub-networks for encoding elevation and land use information surrounding each cell. The network for encoding elevation information comprises four convolutional blocks followed by a fully connected layer for dimensionality reduction and output prediction. The input to the network is a 51×51 single-channel grid. Each convolutional block consists of a 2D convolution layer with a 5×5 kernel and padding of 2 to preserve spatial dimensions, followed by a Leaky ReLU activation ($\alpha = 0.01$) and a 2×2 max-pooling operation (stride $= 2$) to reduce spatial resolution. The number of filters in successive convolutional layers are 16, 32, 16, and 8, respectively. After the four convolutional blocks, the resulting $3 \times 3 \times 8 = 72$ features are flattened and fed into the fully connected layers, yielding a final 3-dimensional output. Land cover data are encoded by the pre-trained ResNet18 [36] network.

DCNN Model. We implement DCNN model as a baseline, which is a deep learning model with the CNN module for feature extraction and multilayer perceptrons (MLP) for $PM_{2.5}$ prediction, as illustrated in figure 7. It takes the concatenation of CNN outputs and other environmental monitoring data and time information as inputs, yielding the final prediction of $PM_{2.5}$ concentrations. There are five MLP layer with hidden sizes of $[512, 256, 64, 16, 1]$, respectively.

DeepAir. For fast inference at scale, we propose DeepAir. It combines the CNN-encoded static features and other meteorological and satellite observation data to train a GBM , as illustrated in figure 8. We use LightGBM as the implementation of GBM, and the CNN structure is identical to the one used in DCNN. The choice of GBM as the back-end regression model balances both efficiency and performance. For one thing, GBM integrates several base learners into a strong predictive model, offering superior fitting capability compared to simpler approaches like linear regression or random forests (as shown in the experiments). For

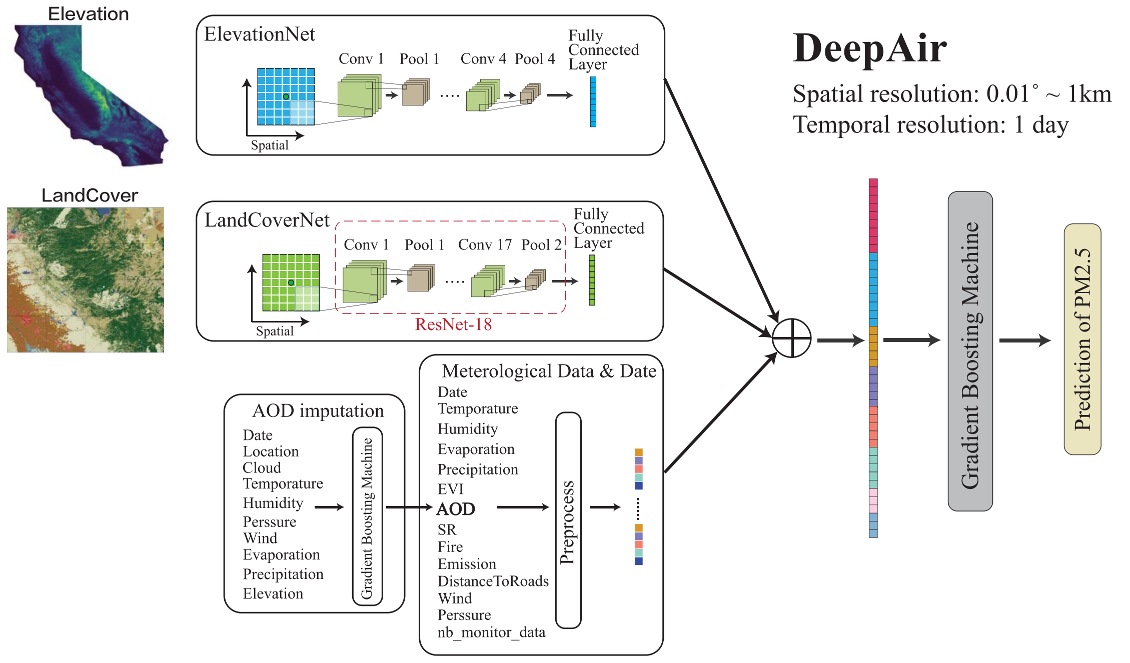

Figure 8. Architecture of the DeepAir model.

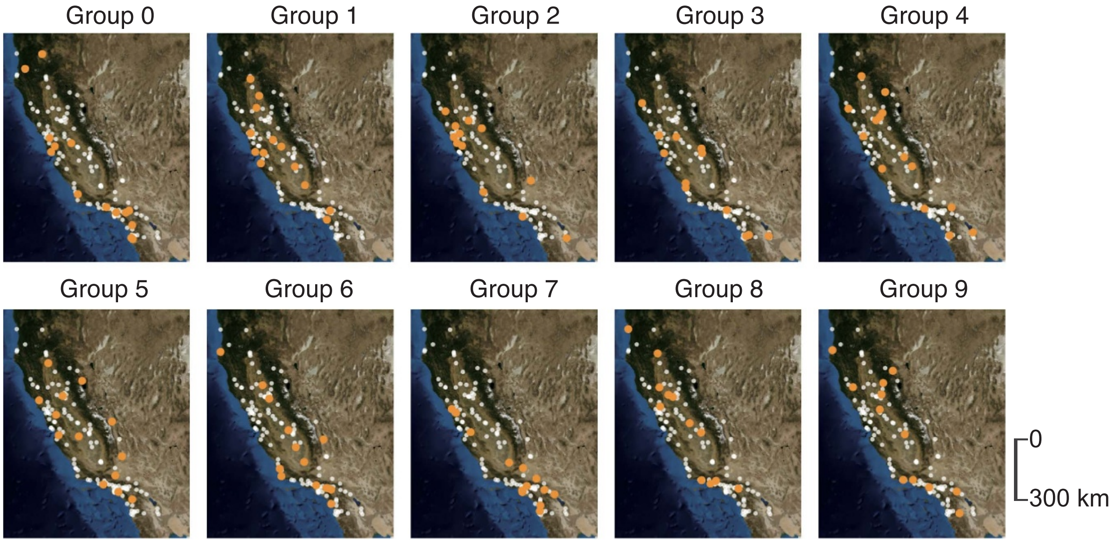

Figure 9. Spatial distribution of different groups of monitors for cross-validation. We split air quality monitors in California into 10 different groups and use a 10-fold validation approach to evaluate the model's performance.

another thing, models such as the support vector machine require considerable computational overhead, which limits their application in large-scale tasks like ours.

3. Experiments

This study has access to daily average concentration records of $PM_{2.5}$ from 130 air quality monitoring stations in California, which are further reorganized as 130 cells with an approximate size of 1 km × 1 km and serves as the dependent variable. Additionally, corresponding meteorological data, satellite observation data, and static data are also available for each cell as the independent variables.

3.1. Settings

In the training phase, we employ a 10-fold cross-validation method, similar to previous studies [37–39]. As shown in figure 9, 130 air quality monitoring stations in California are randomly divided into 10 groups. The

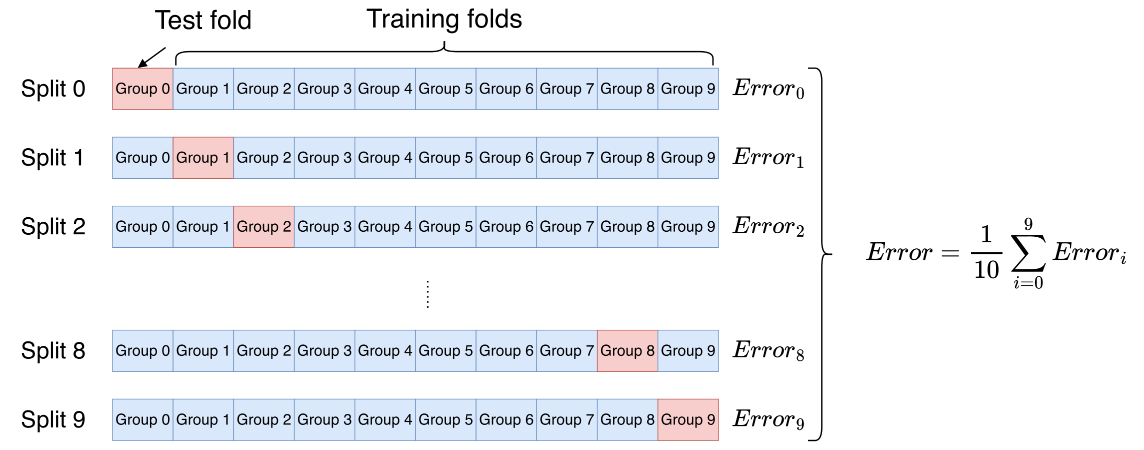
Test fold
Training folds
Split 0
Split 1
Split 2
Split 8
Split 9
Group 0 Group 1 Group 2 Group 3 Group 4 Group 5 Group 6 Group 7 Group 8 Group 9
$Error_0$
$Error_1$
$Error_2$
$Error_8$
$Error_9$
$Error = \frac{1}{10}\sum_{i=0}^{9} Error_i$

Figure 10. Description of the 10-fold cross-validation approach.

Table 1. Performance of different models. We mainly use R^2 to evaluate the model's performance and $\mathrm{MAPE}_{\mathrm{large}}$ focuses on observation data larger than 10 $\mu\mathrm{g\,m}^{-3}$.

Model	R^2	MAE	RMSE	MAPE	$\mathrm{MAPE}_{\mathrm{large}}$	T_{train} (s)	$T_{\mathrm{inference}}$ (s)
RF	0.4374	3.5902	5.7366	0.6569	0.2610	397.27	0.81
GBM	0.5409	3.0152	5.1820	0.5146	0.2468	79.96	2.42
DCNN	0.5278	2.9477	5.2558	0.4556	0.2770	16 741.40	4.83
DeepAir	0.5833	2.7664	4.9369	0.4673	0.2301	13.35	0.21

10-fold cross-validation process is described in figure 10. In each iteration, one group of monitor data are preserved as test data, with the remaining nine groups of data used for training a new model. We compute the evaluation metrics for each division, and the average results of ten independent models are reported to evaluate the performance of different methods.

The primary evaluation metrics are R-squared (R^2), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Additionally, we compute a specific metric called $\mathrm{MAPE}_{\mathrm{large}}$, defined as the MAPE for $\mathrm{PM}_{2.5}$ concentration observations greater than 10 $\mu\mathrm{g\,m}^{-3}$ in the original data. This metric is intended to focus on the model's performance for higher $\mathrm{PM}_{2.5}$ concentrations, as high levels of $\mathrm{PM}_{2.5}$ are particularly impactful for public health and regulatory standards.

Direct comparison with existing $\mathrm{PM}_{2.5}$ prediction models poses challenges due to differences in data sources, spatial resolutions, and temporal granularity. For example, [22] uses random forests(RFs) at a 12 km resolution across the U.S., while [40] employs deep learning for weekly averages at 1 km in California, making alignment impractical. Additionally, the lack of open-source implementations hinders reproducibility. To ensure fairness and consistency, we compare our model with two advanced ML methods and a CNN-based model to our proposed model DeepAir, all tested on the same high-resolution dataset. The evaluated methods are as follows:

- RF: RF Regression Model [19].
- GBM: GBM Regression Model with LightGBM implementation [21].
- DCNN: Deep neural networks with convolutional neural networks.
- DeepAir: convolutional neural networks combined with GBM.

Random forest and GBM model directly utilize preprocessed data in section 2.1.3 as input and conduct the regression task. The architectures of DCNN and DeepAir are introduced in section 2.2. The hyperparameter configurations for each model are included in the appendix.

3.2. Results and analysis

The average results of 10-fold cross-validation are reported in table 1, with detailed results of each fold reported in the appendix. DeepAir achieves the best performance in multiple metrics, including R^2, MAE, RMSE, and $\mathrm{MAPE}_{\mathrm{large}}$. Moreover, the novel combination of a pre-trained CNN and GBM compresses the training and inference overhead, allowing DeepAir to make the fastest inference among all methods.

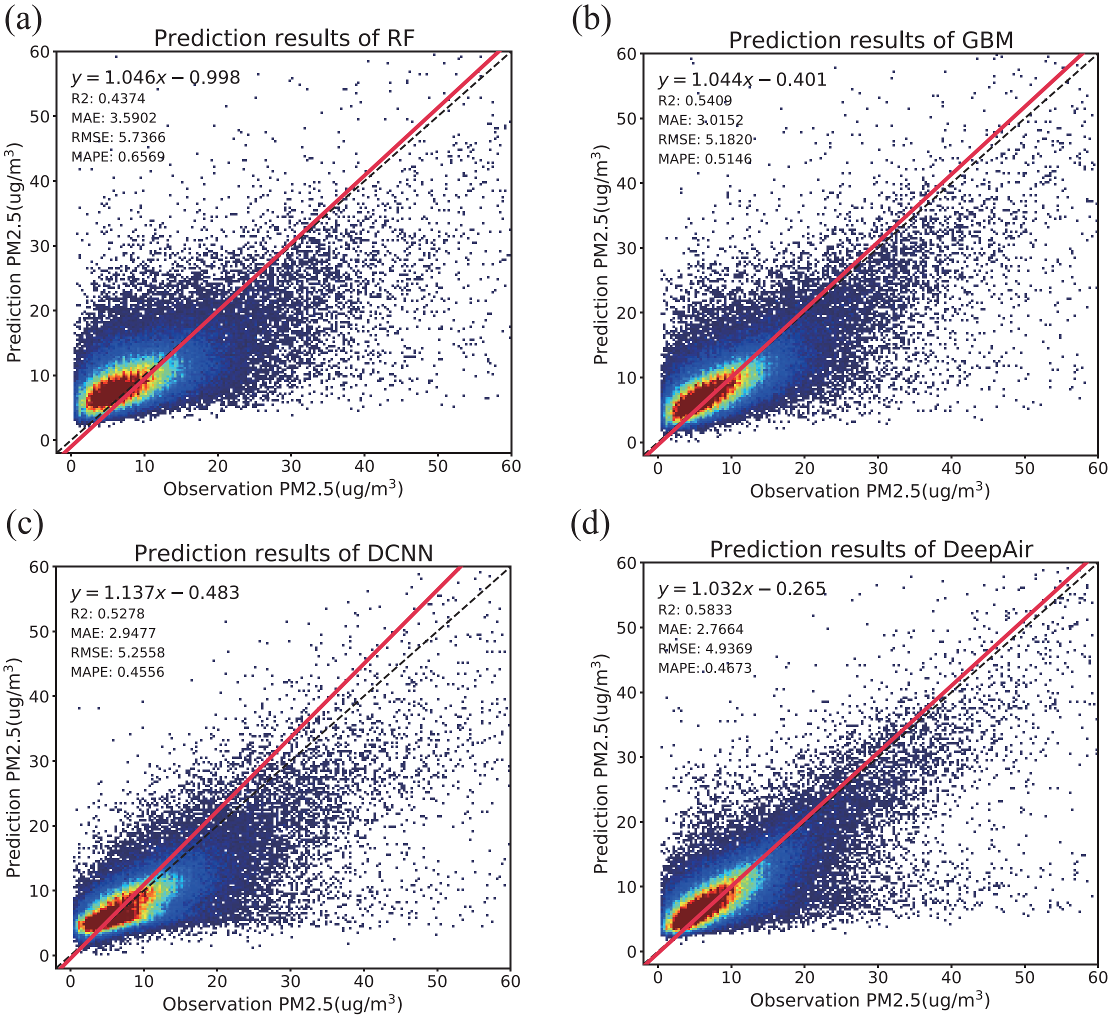

Figure 11. Performance comparison between different models. The dashed line indicates $y = x$, and the red line is a linear equation curve fitted to the actual data from the scatter plot. Experimental results prove that the DeepAir model achieves best R^2.

Comparing GBM and DeepAir, we conclude that the pre-trained CNN module not only accelerates the model by compressing the dimension of static data but also effectively fuses useful features from surrounding areas.

To intuitively compare the spatial inference performance of different models on $\text{PM}_{2.5}$ concentration, we present a comparative graph in figure 11. The x-axis shows actual $\text{PM}_{2.5}$ concentrations, while the y-axis represents model predictions. Using least squares, we fit a linear function (red line) between the ground truth and predicted values for each method, treating ground truth concentrations as independent and predictions as dependent variables. As shown in figure 11, the DeepAir model exhibits the flattest regression slope, indicating predictions closest to actual values. Furthermore, the DeepAir model achieves the highest R^2, demonstrating high accuracy in spatial $\text{PM}_{2.5}$ concentration inference.

Furthermore, we visualize the prediction data distribution of the proposed DeepAir model using violin plots, shown in figure 12. First, we group the test data by month to compare the distribution differences between the model predictions and actual data. Figure 12(a) reveals that the overall predicted distribution aligns well with the observed data, although the model shows a tendency to slightly underestimate low-concentration $\text{PM}_{2.5}$ levels. This can be attributed to our treatment of data below 0.5 $\mu\text{g m}^{-3}$ as outliers, emphasizing higher concentrations that are more relevant to public health.

The original $\text{PM}_{2.5}$ data shows seasonal and monthly trends, with February, March, and April exhibiting more low-concentration data, while December and January show relatively higher concentrations. To further examine these trends, we group the test data by season, as shown in figure 12(b). Here, the $\text{PM}_{2.5}$ observations display a long-tail distribution, with higher concentrations more prevalent in winter. Our model effectively captures these seasonal distribution trends, particularly the variance in high-concentration data during winter. The predicted data closely mirrors the real data distribution across different seasons, demonstrating our model has robust predictive capability in spatial $\text{PM}_{2.5}$ inference.

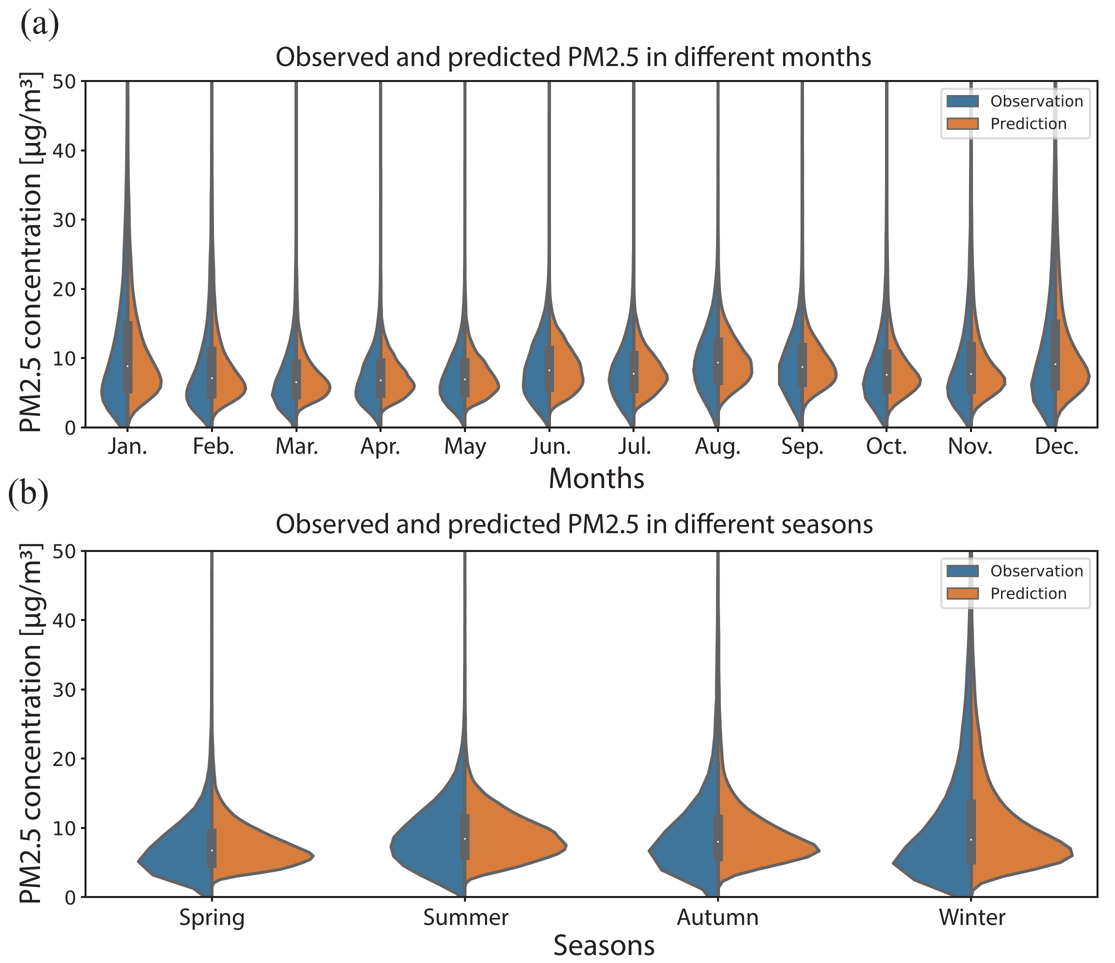

Figure 12. Performance comparison by months and seasons. (a) Comparison of observed and predicted monthly average $PM_{2.5}$ concentrations. (b) Comparison of observed and predicted seasonally average $PM_{2.5}$ concentrations.

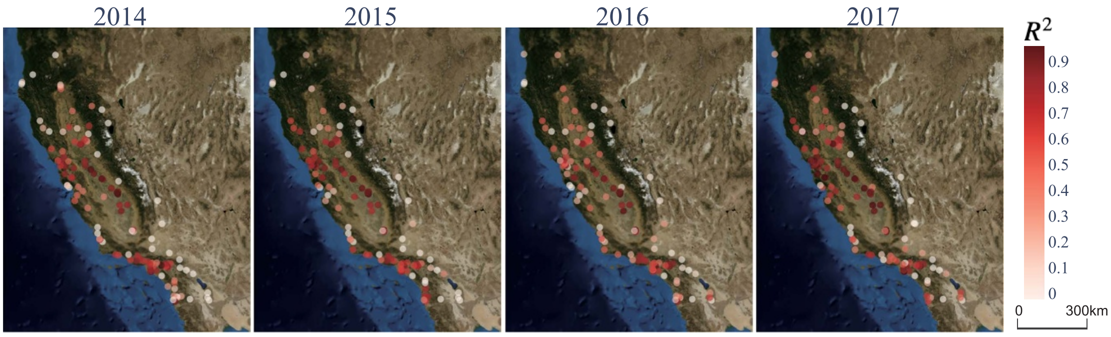

Figure 13. Prediction performance in different years. We calculated the yearly R^2 for every monitor based on its observed $PM_{2.5}$ concentration and the predicted $PM_{2.5}$ concentration generated by our model. Darker colors indicate better predictions on this monitor. The results show that monitors that are close to other monitors usually have higher R^2.

To further examine the influence of nearby air quality monitoring station data on pollutant concentration inference, we group the test data by their associated monitoring stations and record the model's predictions for each group. Next, we categorize the data by year and calculate the R^2 values between the actual and predicted $PM_{2.5}$ concentrations for each monitoring station as a measure of the model's performance across locations and time periods. As presented in figure 13, the R^2 values show little variation in the model's performance on data from the same air quality monitoring stations across different years, suggesting consistent predictive accuracy over time. Spatially, the model achieves higher R^2 values for monitoring stations with dense spatial clustering, indicating that predictions align more closely with actual values in these areas. This outcome highlights the benefits of incorporating data from nearby monitoring

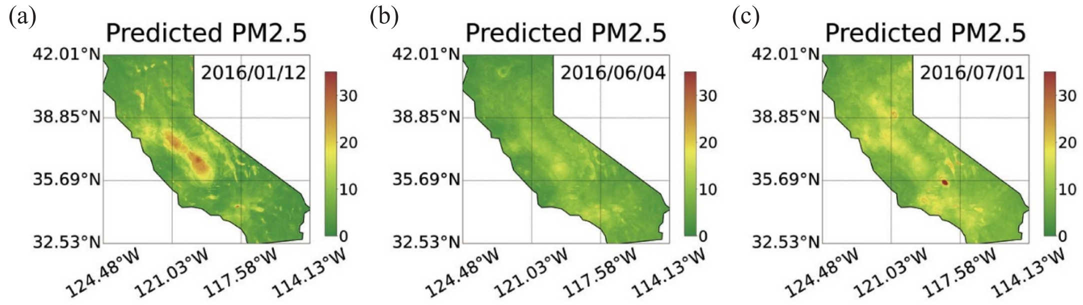

Figure 14. Estimation of $PM_{2.5}$ concentration on typical days Sample days in the California region. (a) Inferred $PM_{2.5}$ concentration for the winter season. (b) Inferred $PM_{2.5}$ concentration for the summer season. (c) Inferred $PM_{2.5}$ concentration during wildfire events.

stations during preprocessing, as pollutant concentrations in adjacent areas tend to be similar. Leveraging spatially consistent information from surrounding monitoring stations enhances the model's ability to infer $PM_{2.5}$ concentration in spatial cells.

3.3. Inferring $PM_{2.5}$ in California

After validating the effectiveness of the DeepAir model, we apply it to estimate the daily average $PM_{2.5}$ concentration for each 1 km × 1 km cell across California from 2014 to 2017. Figure 14 presents visualizations of inferred $PM_{2.5}$ levels on three representative dates-January 12, 2016 (winter), 4 June, 2016 (summer), and 1 July, 2016 (during wildfire activity). As shown in figure 14(a), on 12 January, 2016, typical of winter, the $PM_{2.5}$ concentrations are relatively high. In contrast, on 4 June, 2016, corresponding to summer, the overall $PM_{2.5}$ levels are significantly lower, as illustrated in figure 14(b). For 1 July, 2016, figure 14(c) aligns with the wildfire diffusion simulated in figure 6(b), showing elevated $PM_{2.5}$ levels in areas impacted by wildfire. This pattern demonstrates the model's ability to accurately capture the influence of wildfire events on $PM_{2.5}$ concentration distribution.

4. Possible extensions

The DeepAir framework offers a versatile solution for air pollution-related tasks and can be adapted to predict other pollutants, such as PM_{10}, provided suitable observational data are available. In U.S., PM_{10}, as a criteria pollutant, is observed at ambient monitoring networks similarly as $PM_{2.5}$. However, since PM_{10} originates from different sources compared to $PM_{2.5}$, additional feature engineering would be necessary to tailor the model for such a transfer task. In contrast, applying the methodology to $PM_{0.1}$ is more challenging due to the lack of widespread ambient monitoring data. $PM_{0.1}$ is not regulated as a criteria pollutant by the EPA and thus is not commonly measured at monitoring stations. This limitation underscores the need for alternative data collection strategies, should the prediction of $PM_{0.1}$ be pursued in future research.

5. Conclusion

This paper presents a novel model that combines a pre-trained convolutional neural network with a GBM to infer spatial distributions of $PM_{2.5}$ concentrations. The CNN module encodes static geographic information from surrounding spatial cells, which is then integrated with meteorological data and satellite observations to train a LightGBM model for spatial $PM_{2.5}$ inference. The proposed model demonstrates high accuracy and efficiency in both training and inference. Trained on $PM_{2.5}$ concentration data from 130 air quality monitoring stations across California from 2014 to 2017, the model can be applied to 438 619 1 km × 1 km cells statewide, enabling large-scale, high-resolution inference of daily average $PM_{2.5}$ concentrations. By effectively fusing diverse data sources-including meteorological data, satellite observations, and geographic information-our model provides enhanced spatial $PM_{2.5}$ estimates that supplement existing monitoring station data. These estimates support researchers in analyzing $PM_{2.5}$ distribution patterns and offer a more detailed foundation for air quality assessment and public health applications.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Code availability

The source code is available at https://github.com/humnetlab/DeepAir.

Acknowledgments

This work was supported by the Laboratory Directed Research and Development (LDRD) program of the Lawrence Berkeley National Laboratory. Yanyan Xu was also supported by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the Fundamental Research Funds for the Central Universities. We thank SJTU AI for Science platform for the computing support.

Author contributions

W G, Y X, Z H, J L and M C G conceived the research and designed the analyses. W G, Y X and Z H processed the related data, and performed the analyses. W G, Y X, Z H, and M C G wrote the paper. Y X and M C G supervised the research.

Appendix. Configuration and performance of models

In this section, we present the detailed results of the 10-fold cross-validation. All experiments are conducted on an Intel Xeon Gold 6226 R CPU with 64 cores and NVIDIA GeForce RTX 3090 GPU. The CPU operates on x86-64 architecture and the GPUs run CUDA version 12.1.

A.1. Random forest

Some hyperparameters of the random forest model are specified as follows: the maximum depth of each tree is limited to 256, the random seed is fixed at 0, and the number of trees in the forest is set to 256. Table A1 presents the performance of the random forest model in inferring $PM_{2.5}$ concentration in California.

Table A1. The performance of the Random forest model.

Division	R^2	MAE	RMSE	MAPE	$\text{MAPE}_{\text{large}}$
0	0.4403	3.5295	5.0591	0.6429	0.2497
1	0.6203	3.3890	5.5276	0.5375	0.2516
2	0.3225	3.5323	6.1018	0.5503	0.2910
3	0.3131	3.4390	4.7569	0.7624	0.2685
4	0.5438	3.4021	5.5977	0.6448	0.2381
5	0.1646	3.5438	5.2167	0.8115	0.2650
6	0.4680	4.0264	7.1311	0.5753	0.2739
7	0.3871	3.6178	5.1554	0.6738	0.2698
8	0.4307	3.6082	6.1379	0.6175	0.2282
9	0.3828	3.8241	6.4096	0.7566	0.2682
Average	0.4374	3.5902	5.7366	0.6569	0.2610

A.2. GBM

We use the LightGBM Python package as an implementation of the GBM model, with the following hyperparameters. The random seed is set to 123. The model's evaluation metrics are defined as the MAE and Huber loss. The number of leaves in each tree is set to 512, with a maximum of 512 bins. The maximum depth of the tree is 256, and the learning rate is set to 0.01. During training, the model randomly samples 80% of the features for each iteration and uses 80% of the available data. Bagging occurs every 20 iteration. Training stops if there is no improvement in the validation set score after 100 iterations. Table A2 presents the specific experimental results of the GBM model.

A.3. DCNN

Table A3 presents the experimental results of the DCNN model. Compared to vanilla regression models that only utilize static geographic information of the current grid, the DCNN model employs a CNN structure to

Table A2. The performance of the GBM model.

Division	R^2	MAE	RMSE	MAPE	$\text{MAPE}_{\text{large}}$
0	0.5968	2.8093	4.2937	0.4855	0.2340
1	0.7034	2.8821	4.8858	0.4245	0.2224
2	0.4011	3.1675	5.7370	0.4470	0.2827
3	0.4965	2.8329	4.0723	0.5827	0.2653
4	0.6262	2.8154	5.0668	0.485	0.2375
5	0.2515	3.2141	4.9381	0.6869	0.2731
6	0.5466	3.4697	4.2304	0.4256	0.2677
7	0.5873	2.8207	4.2304	0.5063	0.2221
8	0.5678	2.8487	5.3477	0.4465	0.2199
9	0.4253	3.2632	6.1849	0.6516	0.2462
Average	0.5409	3.0152	5.1820	0.5146	0.2468

extract static geographic information from neighboring regions, generating embedded representations of static geographic features corresponding to the grid. With a substantial number of training instances, the subsequent multilayer perceptron structure can effectively learn the relationships between various input features and the targeted $PM_{2.5}$ concentration. Therefore, this model outperforms vanilla regression models in certain evaluation metrics.

Table A3. The performance of the DCNN model.

Division	R^2	MAE	RMSE	MAPE	$\text{MAPE}_{\text{large}}$
0	0.6170	2.6834	4.1848	0.4188	0.2471
1	0.6750	2.9288	5.1141	0.4009	0.2516
2	0.3238	3.3162	6.0957	0.4204	0.3404
3	0.4741	2.7888	4.1619	0.5457	0.3161
4	0.5270	3.0059	5.6995	0.4613	0.2869
5	0.4431	2.8215	4.2594	0.5492	0.3071
6	0.5587	3.3594	6.4948	0.4214	0.2610
7	0.5705	2.7342	4.3155	0.4302	0.2544
8	0.5667	2.7101	5.3548	0.3639	0.2379
9	0.3803	3.1716	6.4227	0.5619	0.2891
Average	0.5278	2.9477	5.2558	0.4556	0.2770

A.4. DeepAir

Table A4 presents the inference results of the DeepAir model proposed in this paper for the spatial inference of $PM_{2.5}$ concentration in California. The hyperparameters of the LightGBM module are identical to those described in section A.2.

Table A4. The performance of the DeepAir model.

Division	R^2	MAE	RMSE	MAPE	$\text{MAPE}_{\text{large}}$
0	0.6740	2.5397	3.8611	0.4232	0.1981
1	0.7622	2.4659	4.3746	0.3662	0.1958
2	0.4360	2.9089	5.5671	0.4098	0.2599
3	0.4575	2.9454	4.2272	0.5870	0.2820
4	0.6463	2.5707	4.9285	0.4161	0.2242
5	0.4831	2.5305	4.1034	0.5399	0.2347
6	0.4667	3.6061	7.1397	0.5175	0.2542
7	0.6024	2.6602	4.1525	0.4753	0.2189
8	0.7181	2.4191	4.3191	0.3789	0.1937
9	0.4555	3.1483	6.0202	0.5820	0.2659
Average	0.5833	2.7664	4.9369	0.4673	0.2301

ORCID iDs

Wenxuan Guo https://orcid.org/0000-0001-6336-3819
Yanyan Xu https://orcid.org/0000-0001-5429-3177

References

[1] Feng S, Gao D, Liao F, Zhou F and Wang X 2016 The health effects of ambient $PM_{2.5}$ and potential mechanisms *Ecotoxicol. Environ. Saf.* **128** 67–74

[2] Xing Y-F, Xu Y-H, Shi M-H and Lian Y-X 2016 The impact of PM2.5 on the human respiratory system *J. Thoracic Dis.* **8** E69

[3] Wang L, Luo D, Liu X, Zhu J, Wang F, Li B and Li L 2021 Effects of PM2.5 exposure on reproductive system and its mechanisms *Chemosphere* **264** 128436

[4] Thangavel P, Park D and Lee Y-C 2022 Recent insights into particulate matter (PM2.5)-mediated toxicity in humans: an overview *Int. J. Environ. Res. Public Health* **19** 7511

[5] Atkinson R W, Kang S, Anderson H R, Mills I C and Walton H A 2014 Epidemiological time series studies of PM $_{2.5}$ and daily mortality and hospital admissions: a systematic review and meta-analysis *Thorax* **69** 660–5

[6] Wang L *et al* 2018 Taking action on air pollution control in the Beijing-Tianjin-Hebei (BTH) region: progress, challenges and opportunities *Int. J. Environ. Res. Public Health* **15** 306

[7] Xiao Hui C, Dan G, Alamri S and Toghraie D 2023 Greening smart cities: an investigation of the integration of urban natural resources and smart city technologies for promoting environmental sustainability *Sustain. Cities Soc.* **99** 104985

[8] Liu H, Cui W and Zhang Mi 2022 Exploring the causal relationship between urbanization and air pollution: evidence from china *Sustain. Cities Soc.* **80** 103783

[9] He J *et al* 2017 Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities *Environ. Pollut.* **223** 484–96

[10] Buchholz R R *et al* 2021 Air pollution trends measured from terra: Co and aod over industrial, fire-prone and background regions *Remote Sens. Environ.* **256** 112275

[11] Ma Z, Hu X, Sayer A M, Levy R, Zhang Q, Xue Y, Tong S, Bi J, Huang L and Liu Y 2016 Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004–2013 *Environ. Health Perspect.* **124** 184–92

[12] Ma Z, Liu Y, Zhao Q, Liu M, Zhou Y and Bi J 2016 Satellite-derived high resolution $PM_{2.5}$ concentrations in yangtze river delta region of china using improved linear mixed effects model *Atmos. Environ.* **133** 156–64

[13] Grant-Jacob J A and Mills B 2022 Deep learning in airborne particulate matter sensing: a review *J. Phys. Commun.* **6** 122001

[14] Eeftens M *et al* 2012 Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, pm10 and pmcoarse in 20 european study areas; results of the escape project *Environ. Sci. Technol.* **46** 11195–205

[15] Huang L, Zhang C and Bi J 2017 Development of land use regression models for $PM_{2.5}$, SO_2, NO_2 and O_3 in Nanjing, China *Environ. Res.* **158** 542–52

[16] Shi T, Hu Y, Liu M, Li C, Zhang C and Liu C 2020 Land use regression modelling of $PM_{2.5}$ spatial variations in different seasons in urban areas *Sci. Total Environ.* **743** 140744

[17] Liu Y, Paciorek C J and Koutrakis P 2009 Estimating regional spatial and temporal variability of pm2.5 concentrations using satellite data, meteorology and land use information *Environ. Health Perspect.* **117** 886–92

[18] Xiao Q, Geng G, Liang F, Wang X, Lv Z, Lei Y, Huang X, Zhang Q, Liu Y and He K 2020 Changes in spatial patterns of $PM_{2.5}$ pollution in China 2000–2018: impact of clean air policies *Environ. Int.* **141** 105776

[19] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

[20] Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94

[21] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T-Y 2017 Lightgbm: a highly efficient gradient boosting decision tree *Advances in Neural Information Processing Systems* vol 30 (https://doi.org/10.5555/3294996.3295074)

[22] Hu X, Belle J H, Meng X, Wildani A, Waller L A, Strickland M J and Liu Y 2017 Estimating $PM_{2.5}$ concentrations in the conterminous united states using the random forest approach *Environ. Sci. Technol.* **51** 6936–44

[23] Fu D, Xia X, Duan M, Zhang X, Li X, Wang J and Liu J 2018 Mapping nighttime $PM_{2.5}$ from VIIRS DNB using a linear mixed-effect model *Atmos. Environ.* **178** 214–22

[24] Li L, Zhang J, Meng X, Fang Y, Ge Y, Wang J, Wang C, Wu J and Kan H 2018 Estimation of $PM_{2.5}$ concentrations at a high spatiotemporal resolution using constrained mixed-effect bagging models with maiac aerosol optical depth *Remote Sens. Environ.* **217** 573–86

[25] Wang W, Zhao S, Jiao L, Taylor M, Zhang B, Xu G and Hou H 2019 Estimation of $PM_{2.5}$ concentrations in china using a spatial back propagation neural network *Sci. Rep.* **9** 13788

[26] Park Y, Kwon B, Heo J, Hu X, Liu Y and Moon T 2020 Estimating $PM_{2.5}$ concentration of the conterminous united states via interpretable convolutional neural networks *Environ. Pollut.* **256** 113395

[27] Yan X, Zang Z, Jiang Y, Shi W, Guo Y, Li D, Zhao C and Husi L 2021 A spatial-temporal interpretable deep learning model for improving interpretability and predictive accuracy of satellite-based $PM_{2.5}$ *Environ. Pollut.* **273** 116459

[28] Luo Z, Huang F and Liu H 2020 PM2.5 concentration estimation using convolutional neural network and gradient boosting machine *J. Environ. Sci.* **98** 85–93

[29] Zhang L, Na J, Zhu J, Shi Z, Zou C and Yang L 2021 Spatiotemporal causal convolutional network for forecasting hourly PM2.5 concentrations in Beijing, China *Comput. Geosci.* **155** 104869

[30] Zhang K, Yang X, Cao H, Thé J, Tan Z and Yu H 2023 Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning *Environ. Int.* **171** 107691

[31] Thornton P E, Thornton M M, Mayer B W, Wei Y, Devarakonda R, Vose R S and Cook R B 2016 Daymet: daily surface weather data on a 1-km grid for north america, version 3 *ORNL Distributed Active Archive Center* (https://doi.org/10.3334/ornldaac/1328)

[32] Li C and Managi S 2021 Contribution of on-road transportation to $PM_{2.5}$ *Sci. Rep.* **11** 21320

[33] Gately C, Hutyra L R and Wing I S 2019 DARTE annual on-road CO2 Emissions on a 1-km grid, conterminous USA, V2, 1980-2017 *ORNL Distributed Active Archive Center* (https://doi.org/10.3334/ORNLDAAC/1735)

[34] Chen Z-Y, Jin J-Q, Zhang R, Zhang T-H, Chen J-J, Yang J, Ou C-Q and Guo Y 2020 Comparison of different missing-imputation methods for maiac (multiangle implementation of atmospheric correction) aod in estimating daily $PM_{2.5}$ levels *Remote Sens.* **12** 3008

[35] Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, Gu D and Liu Y 2018 Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China plain *Environ. Pollut.* **242** 675–83

[36] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8

[37] Zhai B and Chen J 2018 Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China *Sci. Total Environ.* **635** 644–58

[38] Chen Z-Y, Zhang T-H, Zhang R, Zhu Z-M, Yang J, Chen P-Y, Ou C-Q and Guo Y 2019 Extreme gradient boosting model to estimate $PM_{2.5}$ concentrations with missing-filled satellite data in china *Atmos. Environ.* **202** 180–9

[39] Zamani Joharestani M, Cao C, Ni X, Bashir B and Talebiesfandarani S 2019 $PM_{2.5}$ prediction based on random forest, xgboost and deep learning using multisource remote sensing data *Atmosphere* **10** 373

[40] Li L *et al* 2020 Ensemble-based deep learning for estimating $PM_{2.5}$ over California with multisource big data including wildfire smoke *Environ. Int.* **145** 106143