npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-025-00984-3

Machine learning-guided integration of fixed and mobile sensors for high resolution urban $PM_{2.5}$ mapping

Tianshuai Li[1,3], Xin Huang[3,4], Qingzhu Zhang[1] ✉, Xinfeng Wang[1] ✉, Xianfeng Wang[2] ✉, Anbao Zhu[3], Zhaolin Wei[5], Xinyan Wang[3], Haolin Wang[1], Jiaqi Chen[1], Min Li[1], Qiao Wang[1] & Wenxing Wang[1]

Urban areas exhibit significant gradients in Fine Particulate Matter ($PM_{2.5}$) concentration variability. Understanding the spatiotemporal distribution and formation mechanisms of $PM_{2.5}$ is crucial for public health, environmental justice, and air pollution mitigation strategies. Here, we utilized machine learning and integrated air quality sensor monitoring networks consisting of 200 mobile cruising vehicles and 614 fixed micro–stations to reconstruct $PM_{2.5}$ pollution maps for Jinan's urban area with a high spatiotemporal resolution of 500 m and 1 h. Our study demonstrated that pollution mapping can effectively capture spatiotemporal variations at the urban microscale. By optimizing the spatial design of monitoring networks, we developed a cost-effective air quality monitoring strategy that reduces expenses by nearly 70% while maintaining high precision. The results of multi-model coupling indicated that secondary inorganic aerosols were the primary driving factors for PM2.5 pollution in Jinan. Our work offers a unique perspective on urban air quality monitoring and pollution attribution.

Urban centers are densely populated areas where fine particulate matter ($PM_{2.5}$) pollution has been one of the primary environmental concerns since the 20th century. Prolonged exposure to severe $PM_{2.5}$ pollution poses substantial threats to human health, including increased risks of premature death[1,2]. The emission sources in urban areas are diverse and unevenly distributed. Due to complex physical and chemical processes, $PM_{2.5}$ concentrations exhibit significant local variations over short distances and periods within urban environments[3–5]. High spatiotemporal resolution air quality maps are crucial for capturing fine-scale pollution hotspots, reducing exposure measurement errors, and mitigating public health risks and environmental injustices[6,7]. Moreover, understanding the causes of $PM_{2.5}$ pollution facilitates effective air quality management.

Traditionally, chemical transport model simulations[8–10], land use regression modeling[11], and satellite retrievals[12] have been extensively employed to track the dynamic fluctuations of air quality. However, these methods have inherent limitations when treated with fine-scale urban air pollution. Chemical transport models entail high computational costs and rely on frequent updates of emission inventories. Satellite data, although globally comprehensive, are hindered by issues such as cloud cover[13]. Land use regression models rely on fixed–location monitoring and geographic information system predictor variables, which are constrained by specific local administrative boundaries and often lack the precision required at larger geographic scales[14]. Additionally, sparse monitoring networks prove ineffective at capturing fine-scale pollution hotspots, resulting in inadequate depictions of spatiotemporal heterogeneity of urban air quality[15]. In recent years, advancements in low–cost sensor technology have improved the ability to monitor fine–scale concentration gradients[16–19]. Nevertheless, challenges remain in achieving sufficient data frequency and statistical robustness for temporal representation[20,21]. To effectively map urban air quality and comprehensively understand the spatiotemporal dynamics of air pollution, extensive data collection, and advanced statistical methods are essential.

Generally, intensive emission sources, unfavorable meteorological conditions, and plume transport are key factors that can lead to urban $PM_{2.5}$ pollution events[22–25]. Thorough attribution analysis, utilizing advanced statistical methods, is essential for providing scientific support for regulatory strategies and effective urban air quality management. In recent years, machine learning (ML) has demonstrated remarkable promise in air quality modeling due to its exceptional ability to capture complex and nonlinear relationships between different variables[26–28].

[1]Academician Workstation for Big Data in Ecology and Environment, Environment Research Institute, Shandong University, Qingdao, 266237, China. [2]Shandong Provincial Eco-environment Monitoring Center, Jinan, 250101, China. [3]Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing, 210023, China. [4]Frontiers Science Center for Critical Earth Material Cycling, Nanjing University, Nanjing, 210023, China. [5]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, 430072, China. ✉e-mail: zqz@sdu.edu.cn; xinfengwang@sdu.edu.cn; sdhjwxf@126.com

Compared to traditional methods, ML models can effectively integrate large amounts of multi-source heterogeneous data, such as meteorological, traffic, and geographical information, to make more accurate and real-time predictions of $PM_{2.5}$ pollution events. However, these models often face criticism for their "black–box" nature, which makes it difficult to understand the underlying factors driving their predictions. The development and integration of explainable artificial intelligence (XAI) techniques, such as Shapley additive explanations (SHAP), has become a crucial tool for providing transparency in ML models and elucidating the intricacies of air pollution[29–33]. By quantifying the contributions of individual features, SHAP enables researchers and policymakers to identify key drivers of air pollution events, offering critical insights for crafting targeted mitigation strategies. This interpretability not only enhances the transparency of predictions but also builds trust among stakeholders, facilitating more informed decision–making to improve urban air quality and protect public health.

In this study, we focused on the mapping and attribution of $PM_{2.5}$ pollution in urban Jinan, the capital of a province located in the heavily polluted North China Plain. Utilizing a large–scale, low–cost mobile, and fixed sensor network combined with advanced machine learning algorithms for spatiotemporal modeling, we developed a novel method for generating high spatiotemporal resolution (500 m × 500 m and 1 h) $PM_{2.5}$ datasets. We then explored the optimal arrangement of mobile and fixed sensor monitoring to achieve continuous, high–precision monitoring while minimizing costs, providing valuable insights for urban air quality management. Finally, we accurately quantified the contributions of various factors to urban $PM_{2.5}$ pollution by coupling positive matrix factorization (PMF), the hybrid single-particle Lagrangian integrated trajectory (HYSPLIT), and SHAP, providing a scientific basis for accurately identifying the causes of air pollution and enabling precise control measures.

Results

High spatiotemporal resolution $PM_{2.5}$ pollution mapping

Our dataset can effectively capture the hourly variations in $PM_{2.5}$ concentrations, which is valuable for mitigating the health impacts of acute exposure and facilitating environmental management. Taking a typical severe $PM_{2.5}$ pollution event during February 5 and 6, 2019 (the Spring Festival) in urban Jinan as an example (Fig. 1), our dataset nearly perfectly captured the entire evolution of air quality from clean conditions to severe pollution and subsequent dispersion. From 0:00 to 17:00 (local time) on February 5, data from the atmospheric supersite indicated an hourly average $PM_{2.5\_as}$ concentration of 68 ± 24 μg/m^3. Starting at 18:00, $PM_{2.5}$ pollution began to encroach upon the urban area from the north, initially manifesting as localized contamination. By 0:00 on February 6, severe pollution had enveloped most urban areas, with $PM_{2.5}$ concentrations at the atmospheric supersite peaking at 198 ± 84 μg/m^3. Subsequently, the pollution gradually dissipated as north and northwest winds persisted. This pattern aligns with the occurrence of northwest winds after 23:00 on February 5, as reported by timeanddate (https://www.timeanddate.com/weather/china/jinan/historic?month=2&year=2019). The six China National Environmental Monitoring Center sites within the urban area and Jinan atmospheric supersite also documented fluctuations in $PM_{2.5}$ concentrations throughout the pollution formation and dispersion process between February 5 and 6. These observations are consistent with the spatiotemporal $PM_{2.5}$ pollution mapping results generated by our air quality inference model, as depicted in Supplementary Fig. 1.

Evaluation of mapping before and after reducing fixed micro-stations

Using the same methodology described in the fourth part of "Methods" and integrating data from a reduced number of micro-fixed monitoring sites as

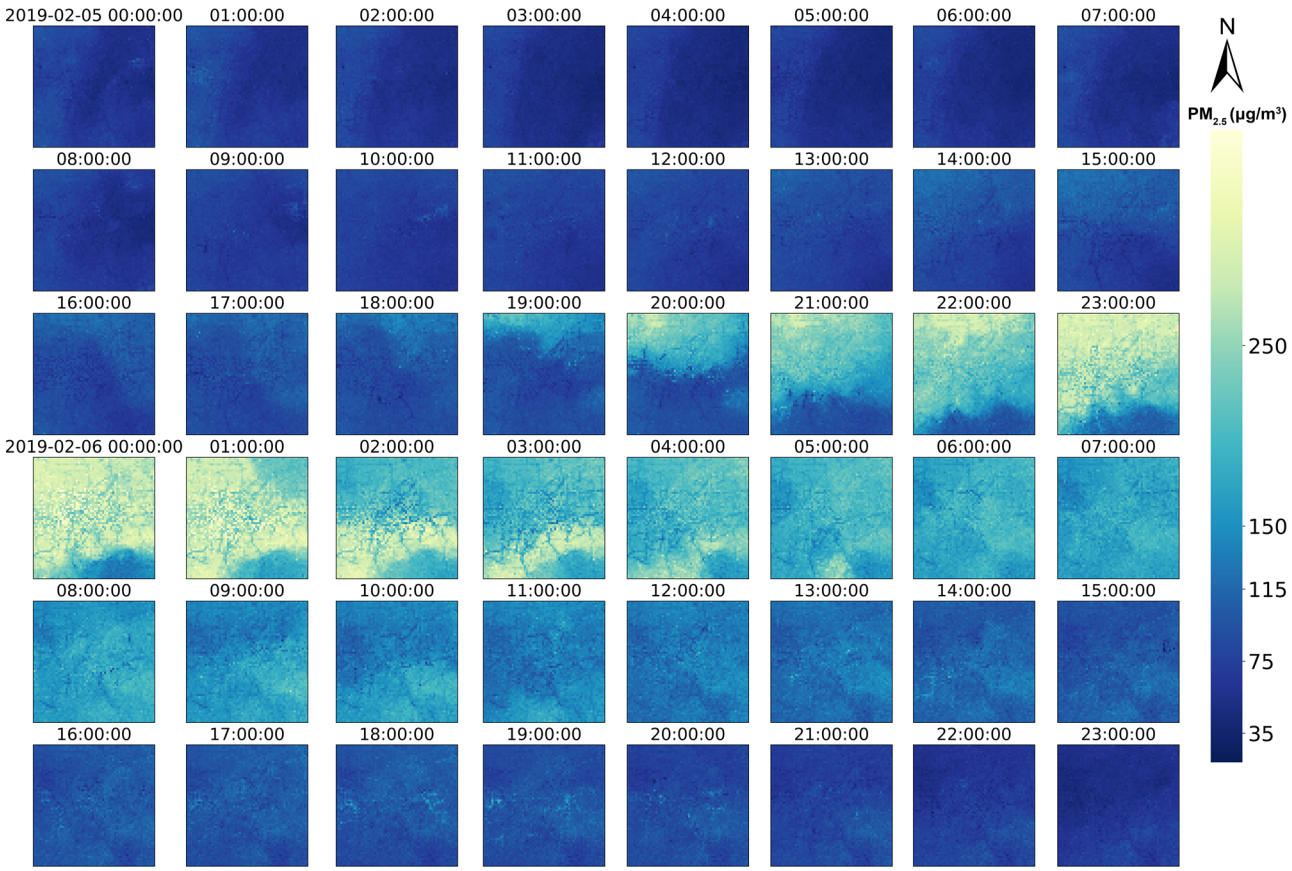

Fig. 1 | Spatiotemporal distribution and evolution of $PM_{2.5}$ concentrations during a pollution event. The selected period of February 5 and 6, 2019, highlights the spatiotemporal dynamics of $PM_{2.5}$ concentrations in the 900 km^2 study area of Jinan, with a 500 m spatial resolution and 1 h temporal resolution.

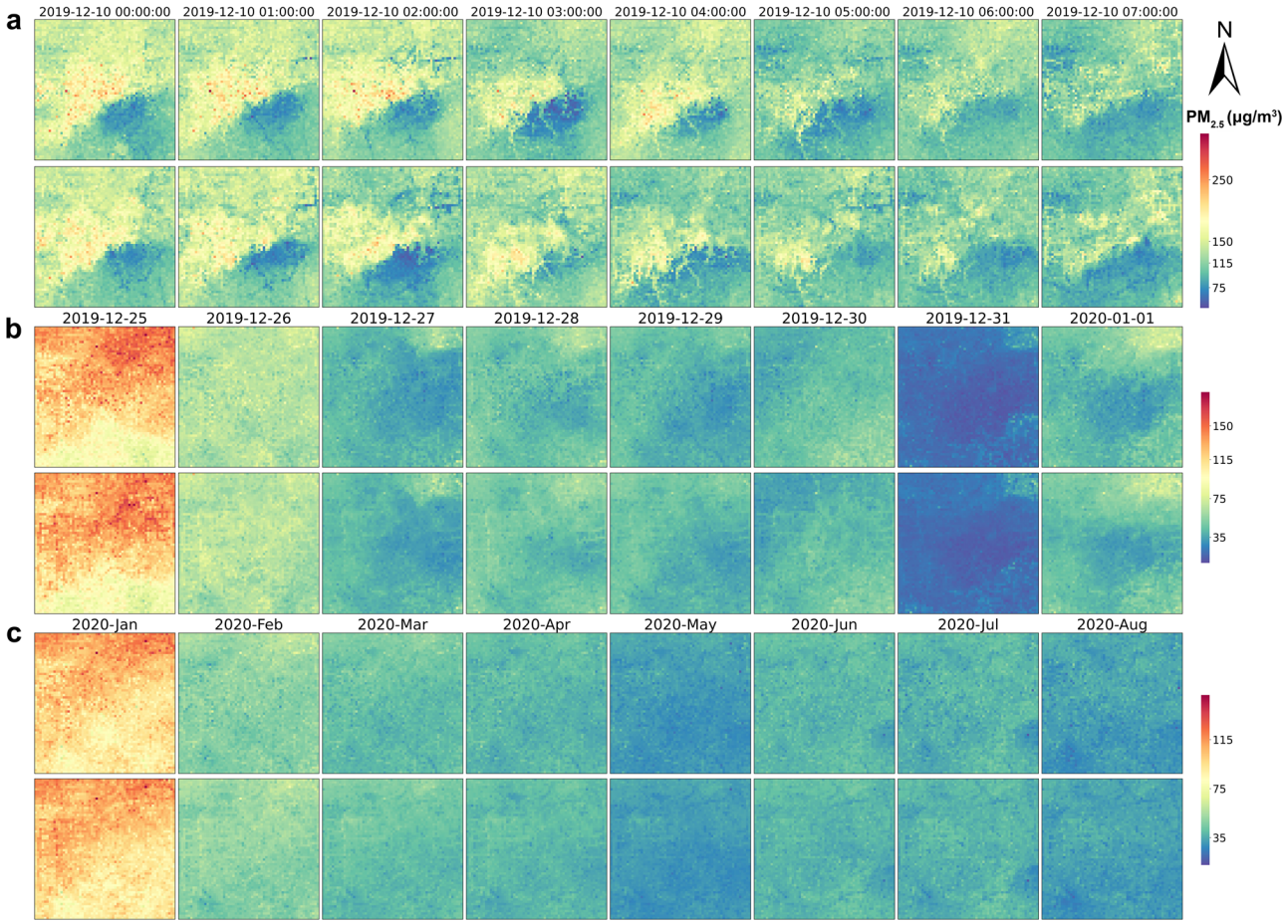

Fig. 2 | The evaluation of the mapping effects on different temporal scales. Comparative evaluation of mapping effects on **a** hourly scale, **b** daily scale, and **c** monthly scale before and after the reduction of micro-fixed monitoring points. The panels in the first, third, and fifth rows display the spatiotemporal distributions of hourly, daily, and monthly averages of $PM_{2.5}$ concentration data products developed from multi-source data before reducing the number of micro-fixed monitoring points. In contrast, the panels in the second, fourth, and sixth rows depict the corresponding distributions after the reduction.

outlined in the fifth part of "Methods", we developed another set of $PM_{2.5}$ data products. This endeavor aims to explore different layouts for combining mobile and micro-fixed monitoring to achieve high-precision urban air quality monitoring while minimizing costs. Figure 2 compares the efficacy of datasets before and after reducing the number of micro-fixed sites on hourly, daily, and monthly scales. We selected a severe $PM_{2.5}$ pollution event on December 10, 2019, to evaluate the effectiveness of both datasets in capturing pollution processes on an hourly scale. By midnight on December 10, $PM_{2.5}$ pollution was severe in most areas of urban Jinan, particularly in the downtown core, with the exception of the southeastern urban area. By 7:00, $PM_{2.5}$ pollution in the downtown area had been alleviated.

On the daily scale, we examined the period from December 25, 2019, to January 1, 2020, as a case study. According to our dataset, on December 25th, most areas exhibited pollution. Pollution gradually abated in the subsequent days, and by December 31st, most parts of the city had achieved clean air status. On the monthly scale, for instance, $PM_{2.5}$ concentrations in January 2020 exceeded that of other listed months in 2020. Seasonal analysis further reveals that winter air quality in Jinan is considerably poorer than in spring, summer, and autumn (Supplementary Fig. 2), likely driven by increased heating emissions and specific meteorological conditions that facilitate pollutant accumulation[34]. Such seasonal disparities highlight winter-specific pollution challenges, underscoring the unique air quality management needs during this season. Overall, across hourly, daily, monthly, and seasonal scales, the patterns of $PM_{2.5}$ concentration variations align with urban area observations (Supplementary Figs. 3-6). Both sets of $PM_{2.5}$ data products effectively capture high-value areas and fine-scale dynamic changes in $PM_{2.5}$ concentrations within urban settings. These findings underscore the significant potential of integrating mobile and micro-fixed monitoring to develop high spatiotemporal $PM_{2.5}$ concentration datasets. Moreover, the dataset built after selectively reducing fixed micro-stations has shown promising results. This outcome not only significantly reduces government expenditures but also holds crucial implications for optimizing the spatial distribution of mobile and fixed micro–station monitoring to enhance the development of high–resolution spatiotemporal datasets.

Key drivers and interpretability attribution of $PM_{2.5}$ in Jinan

Figure 3a displays the driving factors influencing $PM_{2.5}$ concentrations in Jinan, ranked by mean absolute SHAP values. SIA exhibited the highest importance, with an absolute SHAP mean value of 15.83, significantly surpassing other features such as IP (5.70), BB (5.32), T_{as} (4.57), and Dust (3.78). These five drivers predominantly influenced $PM_{2.5}$ concentrations from January 2019 to September 2021. AOD and BLH made comparable contributions, with values ranging from 3 to 3.4, while RH_{as} (2.85) and P (2.52) had slightly weaker impacts. Ox (1.93) played a significant role due to increased radiation and temperature at noon, enhancing photochemical processes[35]. Primary emission sources like CC (1.90) and VE (1.78) contributed less to $PM_{2.5}$ compared to IP, BB, and Dust. Cluster, SSR, WS_{as}, TCC, and WD_{as} had relatively minor predictive effects.

Figure 3b illustrates the response between measured or proxy values of factors and their corresponding SHAP values with respect to $PM_{2.5}$ concentrations, offering additional insights into each factor's influence. Factors such as SIA, IP, BB, and other emission sources showed a pronounced

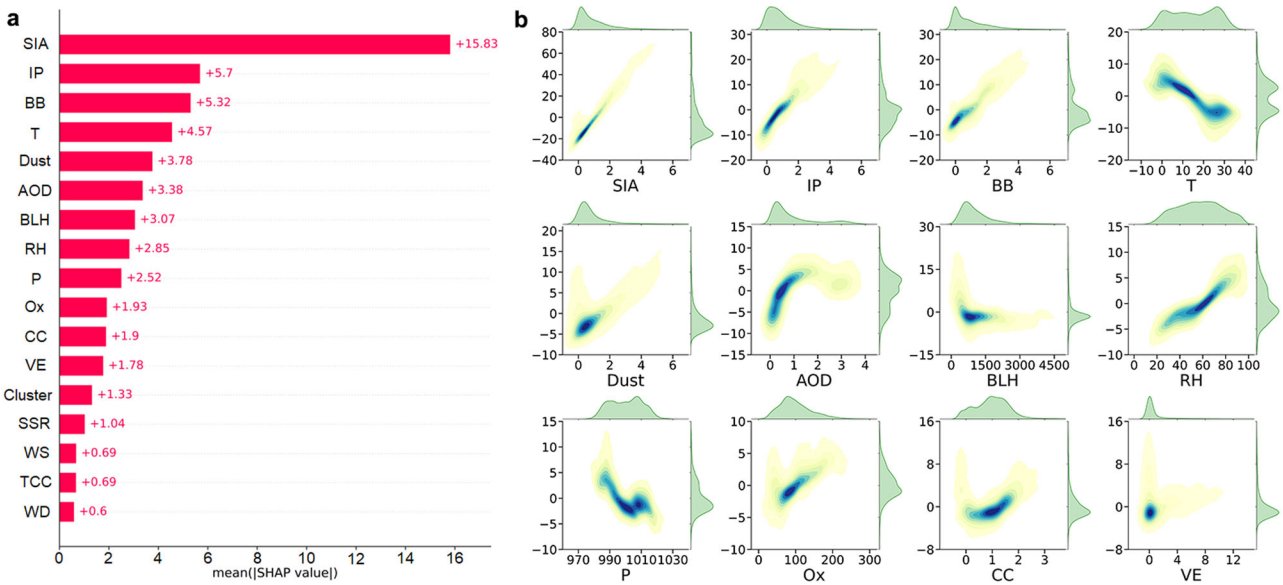

Fig. 3 | Analysis of key driving factors affecting $PM_{2.5}$ concentrations and responses of top predictors in Jinan. The impacts of driving factors on $PM_{2.5}$ concentration in Jinan from January 2019 to September 2021. (**a**) Mean absolute SHAP values of various drivers on $PM_{2.5}$ concentration. (**b**) Responses of SHAP values to the top twelve important predictors for $PM_{2.5}$. Panels are shown as joint plots, where colors in the main plot indicate sample density (dark blue represents high density), with marginal plots showing the distributions of predictor (top) and response (right). SIA: secondary inorganic aerosol; IP: industrial pollution; BB: biomass burning; T: T_{as}, temperature; Dust: dust emission; AOD: aerosol optical depth; BLH: boundary layer height; RH: RH_{as}, relative humidity; P: pressure; O_X: total gaseous oxidant ($NO_2 + O_3$); CC: coal combustion; VE: vehicle emissions; Cluster: air mass trajectory; SSR: surface net solar radiation; WS: WS_{as}, wind speed; TCC: total cloud cover; WD: WD_{as}, wind direction. For more information, see the **Methods** section.

positive contribution to $PM_{2.5}$, highlighting the substantial impact of both secondary formation and primary emissions on air quality in Jinan. Similar positive effects were also observed for AOD, RH_{as}, and O_X. Notably, RH_{as} positively contributed to $PM_{2.5}$ when exceeded 60%, enhancing its formation through aqueous–phase chemistry and hygroscopic growth[29,36,37]. Conversely, P, T_{as}, and BLH exhibited a negative relationship with $PM_{2.5}$ concentrations. Low-pressure systems, usually associated with high humidity, can synergistically enhance $PM_{2.5}$ condensation and coagulation, leading to elevated concentrations[38,39]. Previous literatures have also reported negative effects of atmospheric pressure on $PM_{2.5}$ concentrations in places like Beijing[40] and Hangzhou[39]. A low BLH restricts horizontal and vertical transport, increasing near-surface humidity[41]. High humidity, in turn, enhances aerosol hygroscopic growth, amplifying the positive feedback between aerosols and the boundary layer. This feedback may intensify cross-boundary air pollution transport, exacerbating continuous $PM_{2.5}$ generation[8,42]. Our findings show that SHAP values for air masses 1 and 7 were positive, while air masses 2, 3, and 6 negatively impacted $PM_{2.5}$ concentrations (Supplementary Fig. 7). This suggests that pollutants transported from border areas between northern Zibo and southeastern Binzhou in Shandong, as well as northern Henan and southern Shanxi (Supplementary Fig. 8), increased $PM_{2.5}$ concentrations in Jinan. Conversely, air masses from Mongolia and northeastern Inner Mongolia typically exhibited a cleansing effect. The breakdown in Supplementary Figs. 9 and 10 further delineates the seasonal and pollution-level-specific contributions of various factors to $PM_{2.5}$ concentrations. Seasonal shifts highlight SIA, IP, and BB as key contributors. As $PM_{2.5}$ concentrations rise, emissions from pollution sources and lowered BLH amplify concentrations. Moreover, the impacts of driving factors on the diurnal and nocturnal mechanisms of $PM_{2.5}$ formation are summarized in Fig. 4 and Supplementary Fig. 11.

Discussion

In this study, we employed an advanced light gradient boosting machine (LightGBM), a gradient boosting framework developed by a research team at Microsoft Research Asia[43], to conduct multi-objective model simulations. LightGBM offers lower memory consumption and faster training speed, making it especially suitable for large-scale data analysis, and has demonstrated effective applicability in related studies[44–47]. To validate the performance of LightGBM, we compared it with XGBoost and Random Forest (RF) using 70% and 80% training data splits and various hyperparameter settings. Results in Supplementary Tables 1–3 show that LightGBM consistently outperforms XGBoost and RF, demonstrating robust performance on large datasets. LightGBM's histogram-based decision tree algorithm, which bins continuous features, enables faster node splitting and lower memory usage. In contrast, XGBoost performs precise but computationally intensive splits, while RF requires each tree to be trained independently, increasing resource demands. Additionally, leaf-wise growth and efficient parallelism make LightGBM better suited for data-intensive tasks. Notably, for large datasets, a 70% training set yielded similar or better results than 80%, as more data added minimal benefit while increasing computational costs. The computational efficiency of LightGBM makes it highly applicable to the vast and growing atmospheric environmental datasets. Prioritizing LightGBM in future applications will help meet the demand for high–accuracy, real-time environmental modeling and analysis.

Our study demonstrates that combining mobile and micro-fixed monitoring enables high-resolution air quality monitoring in urban areas, effectively capturing the spatial heterogeneity of $PM_{2.5}$ concentrations. This approach aligns with the needs of policymakers and urban planners who require detailed pollution mapping to make targeted interventions. For example, the ability to dynamically monitor pollution hotspots can inform the timely allocation of resources, such as deploying temporary traffic restrictions or emission reduction measures in critical areas. Furthermore, our optimization of micro-fixed monitoring sites has demonstrated that integrating mobile monitoring maintains coverage while potentially reducing monitoring costs by nearly 70%, from 612 to 184 micro-fixed sites. This reduction illustrates a practical path for cities with limited budgets to still achieve robust air quality monitoring. Additionally, with future research, we aim to evaluate alternative configurations of monitoring networks, such as exploring the effects of different fleet sizes and variations in micro-fixed site density. By assessing these factors, it may be possible to optimize the network further to reduce costs or enhance coverage without sacrificing data quality. These findings can provide a concrete basis for policy recommendations on network design, resource allocation, and budget optimization in

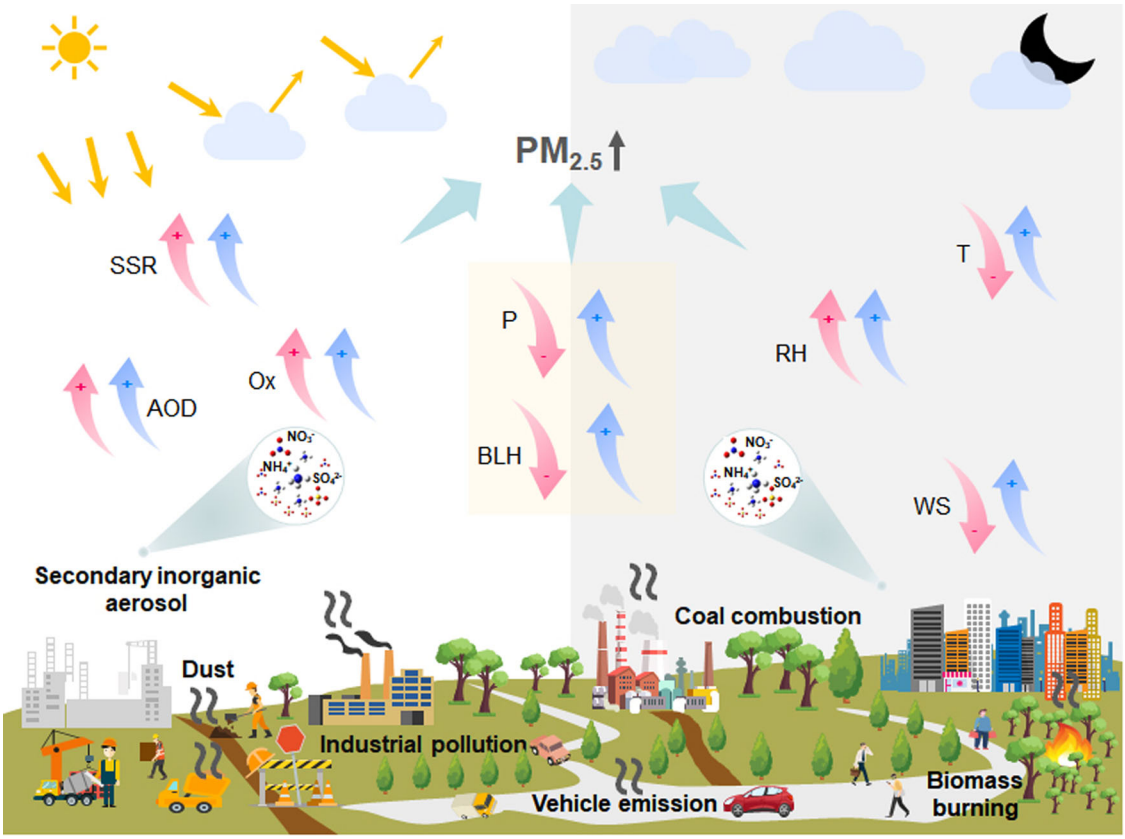

Fig. 4 | Conceptual model depicting daytime and nighttime mechanisms of $PM_{2.5}$ pollution formation in Jinan. Pink and blue arrows denote variations in measured values for each driver and their respective impacts on $PM_{2.5}$ concentrations. The diurnal influences of P and BLH influencing $PM_{2.5}$ concentrations are not observed, and hence, they are depicted in the light orange shaded area centrally in the figure. Relevant calculations are derived from Supplementary Fig. 11.

urban air quality monitoring frameworks. However, the high data volume and computational intensity of our approach underscore the need for efficient data processing infrastructures that can provide actionable, real-time insights to urban planners and policymakers. Our results suggest that strategic reductions in both mobile and micro–fixed monitoring densities may be feasible, enabling more flexible deployment models that balance cost, resource allocation, and monitoring effectiveness. This approach ultimately supports more accessible and economically viable air quality management solutions for urban areas.

Our proposed PMF-HYSPLIT-LightGBM-SHAP coupled model provides a tool for quantifying the contributions of sources, meteorology, and regional transport to $PM_{2.5}$ concentrations, and holds significant potential and value in analyzing the causes of $PM_{2.5}$ pollution. However, due to data limitations, certain constraints still exist. The formation of $PM_{2.5}$ is also influenced by other factors, such as "doming effect" of black carbon[48] and chemical interactions among volatile organic compounds, Ozone (O_3), and $PM_{2.5}$. Further evaluation of their impacts on $PM_{2.5}$ is necessary.

In summary, we propose an innovative method for developing high–resolution spatiotemporal maps of urban $PM_{2.5}$ using mobile and micro-fixed monitoring networks, which can capture fine-scale $PM_{2.5}$ spatial heterogeneity within the urban area. This method not only advances the precision of air quality assessments but also provides valuable insights for optimizing the deployment of mobile and micro-sensor networks, offering significant economic benefits and reference value for deploying monitoring networks in other cities across China and potentially worldwide. Furthermore, by integrating XAI with existing atmospheric models, our research introduces a novel framework for understanding and attributing $PM_{2.5}$ sources, enhancing the depth of analysis and supporting more informed air quality management strategies. Incorporating real-time data into our PMF-HYSPLIT-LightGBM-SHAP coupled model could enhance its dynamic, real-world monitoring capabilities. Integrating continuous measurements from urban monitoring stations, GPS-based traffic data, meteorological inputs, and other diverse data sources through API-based data transfer could enable real-time predictions and model adjustments, providing immediate insights into $PM_{2.5}$ levels. Techniques such as scheduled updates or incremental learning could maintain model stability and accuracy amid continuous data flow. Although real-time monitoring presents challenges, including data latency, quality variability, and processing demands, addressing these could significantly enhance the model's practical utility for quick source identification and adaptive responses to pollution events. Future research will address these technical challenges to further optimize the model for dynamic monitoring scenarios. Looking ahead, we will also focus on refining these models to improve predictive capabilities, expanding the application of our approach to diverse urban environments, and integrating real-time data to dynamically inform and adapt air quality interventions.

Methods

The methodological framework of the study

Figure 5 presents a schematic of the entire build process for this study, illustrating the data sources, machine learning model construction processes, and statistical methods utilized in the analysis. Three models were evaluated: LightGBM, XGBoost, and RF, considering different training data sizes and hyperparameter combinations. The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used as criteria to assess model performance. A description of three performance metrics can be found in Supplementary Text 1. More information about the three model settings can be found in Supplementary Text 2. Based on the comparative results (Supplementary Tables 1–3), a 70% training data split was selected, and the highest-performing LightGBM model was chosen

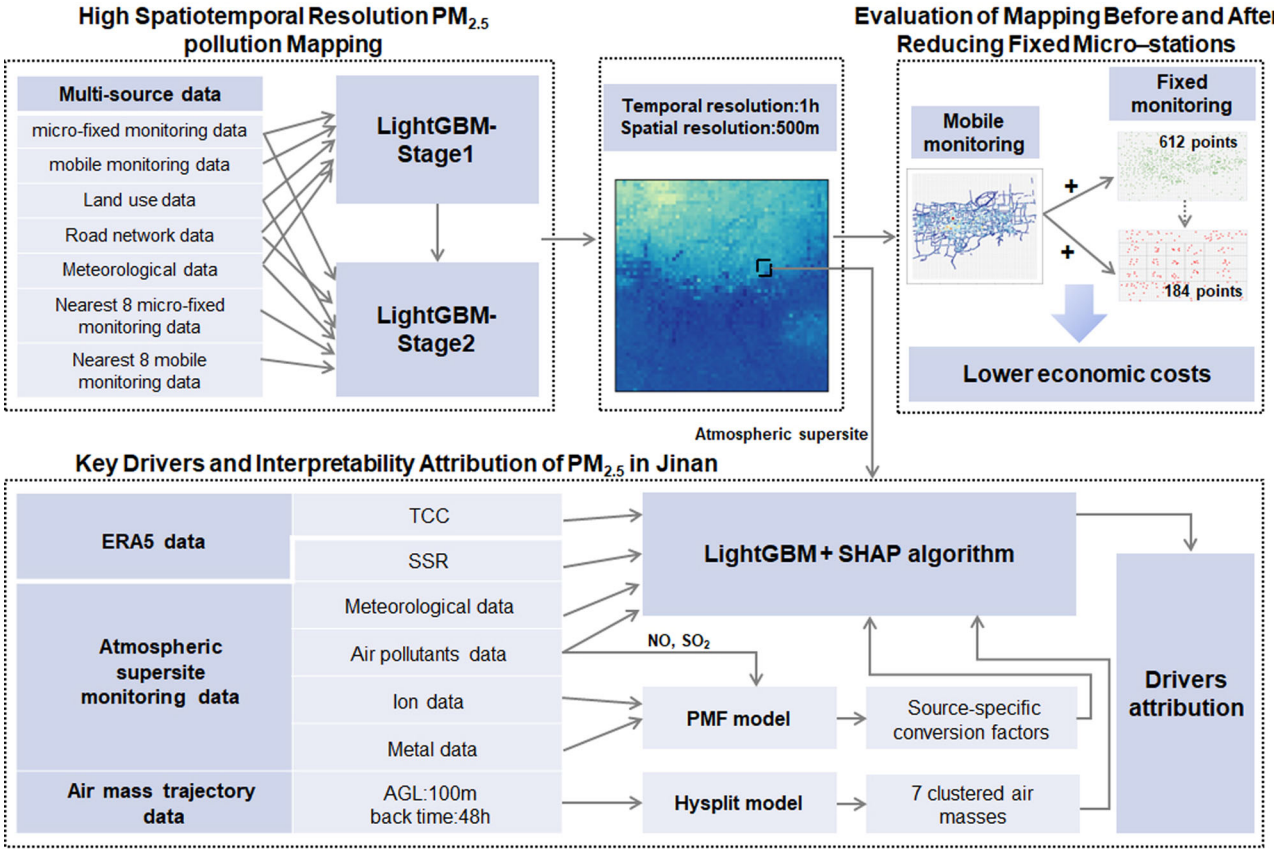

Fig. 5 | Framework diagram of the research development. The methodological framework includes three main components: high spatiotemporal resolution $PM_{2.5}$ pollution mapping, evaluation of mapping before and after reducing fixed micro-stations, and key drivers and interpretability attribution of $PM_{2.5}$ in Jinan.

to conduct high spatiotemporal resolution $PM_{2.5}$ mapping and pollution attribution analysis.

Study area

Jinan, the provincial capital of Shandong Province, with a population exceeding 8 million and over 3.4 million registered vehicles[49], exemplifies a city with a high level of motorization. Considering the operational scope of our pilot vehicles, this study primarily focuses on the urban districts of Jinan (Supplementary Fig. 12). This area spans 900 km^2, is characterized by high population density, and serves as the core region for transportation, administration, commerce, and residential activities in Jinan. This work utilizes nearly three years of continuous air quality data collected from January 1, 2019, to September 13, 2021, through mobile cruising monitoring with vehicle-based sensors and fixed–location monitoring with micro–stations in the study area.

Source of the data

In the context of urban air quality inference and pollution attribution analysis, $PM_{2.5}$ concentrations within a given grid may be influenced by various local factors such as land use characteristics and traffic road networks, as well as external factors, including meteorological information. These factors collectively impact the emission and diffusion of local pollutants and the transport of regional pollutants. Therefore, the collected data comprises seven distinct components: mobile monitoring data ($PM_{2.5\_mobile}$), fixed micro-station monitoring data ($PM_{2.5\_micro\text{-}fixed}$, Wind $speed_{micro\text{-}fixed}$, Wind $direction_{micro\text{-}fixed}$, Relative $humidity_{micro\text{-}fixed}$, $Temperature_{micro\text{-}fixed}$), atmospheric supersite (36.67° N, 117.17°E) data ($PM_{2.5\_as}$, SO_2, NO–NO_2–NO_X, O_3, Wind $speed_{as}$ (WS_{as}), Wind $direction_{as}$ (WD_{as}), Relative $humidity_{as}$ (RH_{as}), $Temperature_{as}$ (T_{as}), Pressure (P), Aerosol Optical Depth (AOD), Boundary Layer Height (BLH), SO_4^{2-}, NH_4^+, NO_3^-, K, Ca, Mn, Zn, Fe, Pb), China National Environmental Monitoring Center (CNEMC) data ($PM_{2.5\_cnemc}$), land use data, road network data, air mass trajectory data, and European Center for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) data (Total cloud cover (TCC), Surface net solar radiation (SSR)). The mobile and fixed micro–station monitoring devices are equipped with four independent particle monitoring sensors that operate synchronously, cross-verify data with each other. In the event of a sensor malfunction, the other sensors continue to function normally, ensuring maximum data reliability. All data were either measured in the field or are publicly available. Detailed information about these multi-source data can be found in Supplementary Table 4.

$PM_{2.5}$ concentration inference model construction

A total of 612 fixed micro-station monitoring are unevenly distributed across 3600 grids in the study area, resulting in many grids lacking continuous monitoring. To address this challenge, we combined multi-source data (mobile monitoring, fixed micro-station monitoring, meteorological, road network, and land use data) to build two-stage LightGBM models to achieve full–coverage continuous monitoring with 500 m × 500 m spatial resolution and 1 h temporal resolution. First, we estimated hourly $PM_{2.5}$ concentrations based on multi-source data for grids with mobile monitoring data but no fixed micro-station data. These estimates served as proxies for the true ground-based fixed monitoring values in these grids. We then used these proxy values to estimate hourly $PM_{2.5}$ concentrations for grids lacking both mobile and fixed micro–station data, enabling high spatiotemporal resolution mapping of $PM_{2.5}$ concentrations across the entire study area.

Specifically, meteorological, road network, land use, mobile monitoring, and fixed micro-station monitoring data were allocated to grids. We used meteorological, land use, and mobile monitoring data from grids with both mobile and fixed micro-station monitoring data as input variables, while fixed micro-station data served as labels for constructing $LightGBM_{reconstruct}$-Stage1. Subsequently, this model estimated hourly

$PM_{2.5}$ concentrations in grids with mobile monitoring data but without fixed micro-station data, with these estimates serving as proxies for ground-based fixed monitoring values. $LightGBM_{reconstruct}$-Stage2 was trained using the nearest spatially adjacent data points, including eight fixed micro-station observations, eight mobile observations, and meteorological, land use, and road network data from the same hour. These variables, along with $PM_{2.5}$ concentrations from grids equipped with fixed micro-station monitoring data, were used to predict hourly $PM_{2.5}$ concentrations in grids lacking both mobile and fixed micro-station monitoring data. Figure 5 and Supplementary Fig. 13 show schematic diagrams of the construction processes, while Supplementary Text 3 presents a more detailed description.

Both $LightGBM_{reconstruct}$ models performed well on the testing set (Supplementary Fig. 14a, b), with R^2 of 0.91 for the first stage model ($LightGBM_{reconstruct}$-Stage1) and R^2 of 0.97 for the second stage model ($LightGBM_{reconstruct}$-Stage2), indicating that constructed models effectively infer $PM_{2.5}$ concentrations.

Inference model development based on new strategy

In an effort to optimize the budget while maintaining effective air quality monitoring, we divided the study area into 17 target regions as shown in Fig. 6. Following the algorithm outlined below, we selected the six nearest fixed micro-stations to the center point of each target region, ranging from target region 1 to target region 17. The specific algorithm formulas are as follows:

$$D_i = \sqrt{\left(Lng_i - \text{center\_lng}\right)^2 + \left(Lat_i - \text{center\_lat}\right)^2} \quad (1)$$

$$X = \left\{\left(Lng_i, Lat_i, D_i\right), |, i = 1, 2, \ldots, n\right\} \quad (2)$$

$$H = H_6(S(D, X)) \quad (3)$$

Lng_i and Lat_i denote the longitude and latitude of fixed micro-stations in the target regions, respectively. Center_lon and Center_lat refer to the longitude and latitude of each target region's center point. D_i represents the distance from each fixed micro-station to the center. X represents the set of distances from each fixed micro-station to the center point, and N signifies the number of the fixed micro-station in the target areas. Here, we defined the sorting operation function S and selection function H. Therefore, H_6 select the 6 fixed micro–stations closest to the center of each target region based on the smallest sorted distances.

Based on the above process, we selected 184 out of 612 fixed micro-stations and replicated the steps above-mentioned in the fourth part of "Methods" to train the $LightGBM_{strategy}$ model for inferring $PM_{2.5}$ concentrations. Surprisingly, the two–stage $LightGBM_{strategy}$ model still performed strongly, achieving R^2 values of 0.87 and 0.97, respectively (Supplementary Fig. 14c,d).

Development of $PM_{2.5}$ pollution attribution model

To identify key factors influencing $PM_{2.5}$ concentrations, we integrated the PMF model, HYSPLIT model, LightGBM model, and SHAP algorithm to establish an interpretable predictive model and elucidate each feature's contribution. The main equation for SHAP is:

$$f(x_i) = \varphi_q(f, x) + \sum_{p=1}^{K} \varphi_p(f, x_i) \quad (4)$$

Here, $f(x_i)$ represents the predicted value for each sample (x_i) with K features and $\varphi_q(f, x)$ is the output expectation of the model for all samples, and $\varphi_p(f, x_i)$ denotes the Shapley value of feature p on the predicted outcome of the sample (x_i). Detailed calculations are provided in Supplementary Text 4.

Initially, three species of water-soluble ions (NH_4^+, NO_3^-, SO_4^{2-}), metal elements (K, Ca, Mn, Fe, Zn, Pb), and atmospheric pollutants (NO, SO_2) were selected for PMF analysis using US EPA PMF v5.0. We identified six sources: Coal Combustion (CC), Dust Emissions (Dust), Industrial

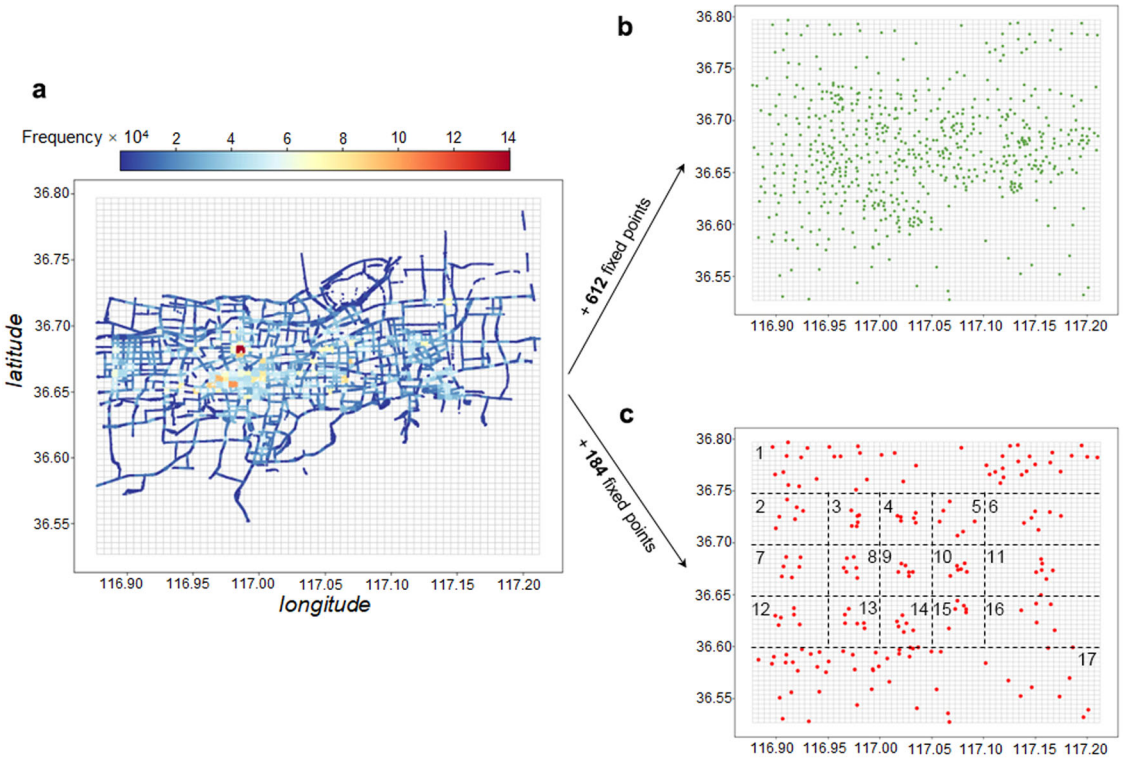

Fig. 6 | Distribution of mobile monitoring driving routes and micro-fixed monitoring sites. a The heatmap of mobile monitoring driving routes. **b** Original distribution of 612 micro-fixed monitoring sites. **c** The distribution of 184 micro-fixed monitoring sites retained using the new strategy.

Pollution (IP), Vehicular Emissions (VE), Biomass Burning (BB), and Secondary Inorganic Aerosol (SIA). Further details about PMF analysis can be found in Supplementary Text 5 and Supplementary Fig. 15. Subsequently, the HYSPLIT model was applied to calculate air clusters and characterize the air mass transport. For each hourly measurement, 48 h backward air mass trajectories at 100 m above ground level were calculated and clustered into seven groups (Supplementary Fig. 8). Total gaseous oxidant ($O_X = NO_2 + O_3$) served as a proxy for atmospheric photochemical oxidation conditions. We then input multiple variables including six emission sources (SIA, CC, BB, IP, Dust, VE), regional transport characteristics (Clusters), atmospheric oxidation condition (O_X), and meteorological parameters (BLH, WS_{as}, T_{as}, RH_{as}, AOD, TCC, SSR, P, WD_{as}) into the $LightGBM_{cause}$ model to analyze $PM_{2.5}$ attribution. The $LightGBM_{cause}$ model achieved strong performance with an R^2 of 0.89, RMSE of 14.18, and MAE of 8.93 (Supplementary Fig. 16), demonstrating an effective explanation of $PM_{2.5}$ concentration changes. Finally, we employed the SHAP algorithm as an XAI tool to quantify each feature's contribution, thereby elucidating the factors contributing to $PM_{2.5}$ pollution in Jinan.

Code availability

Codes generated during this study are available from the corresponding author upon request. Our high spatiotemporal resolution $PM_{2.5}$ dataset for Jinan is openly accessible at https://zenodo.org/records/14961689 for public use.

Received: 31 August 2024; Accepted: 27 February 2025;
Published online: 10 March 2025

References

1. Kaufman, J. D. et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. *Lancet* **388**, 696–704 (2016).
2. Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M. & Pope, C. A. I. Ambient $PM_{2.5}$ reduces global and regional life expectancy. *Environ. Sci. Technol. Lett.* **5**, 546–551 (2018).
3. Wen, Y. et al. Dynamic traffic data in machine-learning air quality mapping improves environmental justice assessment. *Environ. Sci. Technol.* **58**, 3118–3128 (2024).
4. Caubel, J. J., Cados, T. E., Preble, C. V. & Kirchstetter, T. W. A distributed network of 100 black carbon sensors for 100 days of air quality monitoring in West Oakland, California. *Environ. Sci. Technol.* **53**, 7564–7573 (2019).
5. Apte, J. S. et al. High-resolution air pollution mapping with Google Street View cars: exploiting big data. *Environ. Sci. Technol.* **51**, 6999–7008 (2017).
6. Wang, Y. et al. Location-specific strategies for eliminating US national racial-ethnic $PM_{2.5}$ exposure inequality. *Proc. Natl Acad. Sci. USA* **119**, e2205548119 (2022).
7. Alexeeff, S. E. et al. High-resolution mapping of traffic related air pollution with Google Street View cars and incidence of cardiovascular events within neighborhoods in Oakland, CA. *Environ. Health* **17**, 38 (2018).
8. Huang, X. et al. Amplified transboundary transport of haze by aerosol–boundary layer interaction in China. *Nat. Geosci.* **13**, 428–434 (2020).
9. Qin, Y. et al. Amplified positive effects on air quality, health, and renewable energy under China's carbon neutral target. *Nat. Geosci.* **17**, 411–418 (2024).
10. Huang, X. et al. Escalating Wildfires in Siberia driven by climate feedbacks under a warming arctic in the 21st Century. *AGU Adv.* **5**, e2023AV001151 (2024).
11. Messier, K. P. et al. Mapping air pollution with Google Street View cars: efficient approaches with mobile monitoring and land use regression. *Environ. Sci. Technol.* **52**, 12563–12572 (2018).
12. Wei, J. et al. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sens. Environ.* **252**, 112136 (2021).
13. Christiansen, A. E., Carlton, A. G. & Henderson, B. H. Differences in fine particle chemical composition on clear and cloudy days. *Atmos. Chem. Phys.* **20**, 11607–11624 (2020).
14. Eeftens, M. et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).
15. Minet, L. et al. Development and comparison of air pollution exposure surfaces derived from on-road mobile monitoring and short-term stationary sidewalk measurements. *Environ. Sci. Technol.* **52**, 3512–3519 (2018).
16. Liu, X. et al. Spatiotemporal characteristics and driving factors of black carbon in Augsburg, Germany: Combination of mobile monitoring and street view images. *Environ. Sci. Technol.* **55**, 160–168 (2021).
17. Lloyd, M. et al. Predicting within-city spatial variations in outdoor ultrafine particle and black carbon concentrations in Bucaramanga, Colombia: a hybrid approach using open-source geographic data and digital images. *Environ. Sci. Technol.* **55**, 12483–12492 (2021).
18. Zhao, B. et al. Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning. *Environ. Sci. Technol.* **55**, 5579–5588 (2021).
19. Wang, Y. et al. Vehicular ammonia emissions significantly contribute to urban $PM_{2.5}$ pollution in two Chinese megacities. *Environ. Sci. Technol.* **57**, 2698–2705 (2023).
20. Wang, A. et al. Key themes, trends, and drivers of mobile ambient air quality monitoring: a systematic review and meta-analysis. *Environ. Sci. Technol.* **57**, 9427–9444 (2023).
21. Apte, J. S. & Manchanda, C. High-resolution urban air pollution mapping. *Science* **385**, 380–385 (2024).
22. Zhai, S. et al. Fine particulate matter ($PM_{2.5}$) trends in China, 2013–2018: separating contributions from anthropogenic emissions and meteorology. *Atmos. Chem. Phys.* **19**, 11031–11041 (2019).
23. Li, Z. et al. Aerosol and boundary-layer interactions and impact on air quality. *Natl Sci. Rev.* **4**, 810–833 (2017).
24. Zheng, G. J. et al. Exploring the severe winter haze in Beijing: the impact of synoptic weather, regional transport and heterogeneous reactions. *Atmos. Chem. Phys.* **15**, 2969–2983 (2015).
25. Wang, G. et al. Persistent sulfate formation from London Fog to Chinese haze. *Proc. Natl Acad. Sci. USA* **113**, 13630–13635 (2016).
26. Xu, P. et al. Fertilizer management for global ammonia emission reduction. *Nature* **626**, 792–798 (2024).
27. Luo, K., Wang, X., de Jong, M. & Flannigan, M. Drought triggers and sustains overnight fires in North America. *Nature* **627**, 321–327 (2024).
28. Gerges, F., Llaguno-Munitxa, M., Zondlo, M. A., Boufadel, M. C. & Bou-Zeid, E. Weather and the city: machine learning for predicting and attributing fine scale air quality to meteorological and urban determinants. *Environ. Sci. Technol.* **58**, 6313–6325 (2024).
29. Hou, L. et al. Revealing drivers of haze pollution by explainable machine learning. *Environ. Sci. Technol. Lett.* **9**, 112–119 (2022).
30. Li, T. et al. Contributions of various driving factors to air pollution events: Interpretability analysis from machine learning perspective. *Environ. Int.* **173**, 107861 (2023).
31. Li, T. et al. Characteristics of secondary inorganic aerosols and contributions to $PM_{2.5}$ pollution based on machine learning approach in Shandong Province. *Environ. Pollut.* **337**, 122612 (2023).
32. Lee, Y. et al. Unveiling teleconnection drivers for heatwave prediction in South Korea using explainable artificial intelligence. *Npj Clim. Atmos. Sci.* **7**, 1–12 (2024).
33. Pernov, J. B. et al. Pan-arctic methanesulfonic acid aerosol: source regions, atmospheric drivers, and future projections. *Npj Clim. Atmos. Sci.* **7**, 1–18 (2024).

34. Lv, Z., Wei, W., Cheng, S., Han, X. & Wang, X. Meteorological characteristics within boundary layer and its influence on $PM_{2.5}$ pollution in six cities of North China based on WRF. *Chem. Atmos. Environ.* **228**, 117417 (2020).
35. Fu, X. et al. Persistent heavy winter nitrate pollution driven by increased photochemical oxidants in Northern China. *Environ. Sci. Technol.* **54**, 3881–3889 (2020).
36. Le, T. et al. Unexpected air pollution with marked emission reductions during the COVID-19 outbreak in China. *Science* **369**, 702–706 (2020).
37. Wang, H. et al. High N_2O_5 concentrations observed in urban Beijing: implications of a large nitrate formation pathway. *Environ. Sci. Technol. Lett.* **4**, 416–420 (2017).
38. Chen, Z. et al. Influence of meteorological conditions on $PM_{2.5}$ concentrations across China: a review of methodology and mechanism. *Environ. Int.* **139**, 105558 (2020).
39. Jian, L., Zhao, Y., Zhu, Y.-P., Zhang, M.-B. & Bertolatti, D. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Sci. Total Environ.* **426**, 336–345 (2012).
40. Zhang, H., Wang, Y., Hu, J., Ying, Q. & Hu, X.-M. Relationships between meteorological parameters and criteria air pollutants in three megacities in China. *Environ. Res.* **140**, 242–254 (2015).
41. Tie, X. et al. Severe pollution in China amplified by atmospheric moisture. *Sci. Rep.* **7**, 15760 (2017).
42. Peng, J. et al. Explosive secondary aerosol formation during severe haze in the North China Plain. *Environ. Sci. Technol.* **55**, 2189–2207 (2021).
43. Ke, G. et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst* **30**, 3146–3154 (2017).
44. Wei, J. et al. Himawari-8-derived diurnal variations in ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). *Atmos. Chem. Phys.* **21**, 7863–7880 (2021).
45. Wang, S. et al. Diagnosing drivers of $PM_{2.5}$ simulation biases in China from meteorology, chemical composition, and emission sources using an efficient machine learning method. *Geosci. Model Dev.* **17**, 3617–3629 (2024).
46. Bashan, N. F., Li, W. & Wang, Q. R. Dynamics of $PM_{2.5}$ and network activity during extreme pollution events. *Npj Clim. Atmos. Sci.* **7**, 1–8 (2024).
47. Wang, S. et al. Reconstructing long-term (1980–2022) daily ground particulate matter concentrations in India (LongPMInd). *Earth Syst. Sci. Data* **16**, 3565–3577 (2024).
48. Ding, A. J. et al. Enhanced haze pollution by black carbon in megacities in China. *Geophys. Res. Lett.* **43**, 2873–2879 (2016).
49. Jinan Municipal Bureau of Statistics. Jinan Statistical Yearbook (2022).

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (22236004 and 42361144721) and the Taishan Scholars Foundation of Shandong Province (ts201712003).

Author contributions

T.L. conducted the data analyses and model simulations and wrote the paper. X.H. supervised the paper. Q.Z., X.W., and X.W. conceived and supervised the paper. Z.W. provided the support of model simulations. A.Z., X.W., H.W., J.C., and M.L. participated in the data acquisition and statistical analysis of the paper. Q.W. and W.W. assisted in supervising the manuscript. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-00984-3.

Correspondence and requests for materials should be addressed to Qingzhu Zhang, Xinfeng Wang or Xianfeng Wang.