A Novel Method for Quantifying the Contribution of Regional Transport to $PM_{2.5}$ in Beijing (2013-2020): Combining Machine Learning with Concentration-Weighted Trajectory Analysis

Kang Hu[1], Hong Liao[1], Dantong Liu[2], Jianbing Jin[1], Lei Chen[1], Siyuan Li[2], Yangzhou Wu[3], Changhao Wu[4], Shitong Zhao[2], Xiaotong Jiang[5], Ping Tian[6,7], Kai Bi[6,7], Ye Wang[8], Delong Zhao[6,7]

[1]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, Nanjing University of Information Science & Technology, Nanjing 210044, China.

[2]Department of Atmospheric Sciences, School of Earth Sciences, Zhejiang University, Hangzhou 310058, China.

[3]Guangxi Key Laboratory of Environmental Pollution Control Theory and Technology, Guilin University of Technology, Guilin 541004, China.

[4]Institute of International Rivers and Eco-security, Yunnan University, Kunming 650091, China.

[5]College of Biological and Environmental Engineering, Shandong University of Aeronautics, Binzhou, 256600, China.

[6]Beijing Key Laboratory of Cloud, Precipitation and Atmospheric Water Resources, Beijing 100089, China.

[7]Field Experiment Base of Cloud and Precipitation Research in North China, China Meteorological Administration, Beijing 100089, China.

[8]Key Laboratory of Meteorological Disaster, Ministry of Education (KLME)/Joint International Research Laboratory of Climate and Environment Change (ILCEC)/ Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science and Technology, Nanjing 210044, China.

Corresponding author: Hong Liao (hongliao@nuist.edu.cn)

Abstract

Fine particulate matter ($PM_{2.5}$) is closely linked to human health, with its sources generally divided into local emissions and regional transport. This study combined concentration-weighted trajectory (CWT) analysis with the HYSPLIT trajectory ensemble to obtain hourly-resolution pollutant source results. The Extreme Gradient Boosting (XGBoost) model was then employed to simulate local emissions and ambient $PM_{2.5}$ in Beijing from 2013 to 2020. The results revealed that clean air masses influencing the Beijing area mainly originated from the north and east regions, exhibiting a strong winter and weak summer pattern. Following the implementation of the Air Pollution Prevention and Control Action Plan (Action Plan) by the Chinese government in 2017, pollution in Beijing decreased significantly, with the most substantial reduction in regional transport pollution events occurring in the west region during summer. Regional transport pollution events were most frequent in spring, up to 1.8 times higher than in winter. Pollutants mainly originated from the west and south regions, while polluted air masses from the east showed the least reduction, and the proportion of pollution sources from this region was gradually increasing. From 2013 to 2020, local emissions were the main contributors to pollution events in Beijing. The Action Plan has more effectively reduced pollution caused by regional transport, particularly during autumn and winter. This finding underscores the importance of Beijing prioritizing local emission reduction while also considering potential contributions from the east region to effectively mitigate pollution events.

Keywords: Fine particulate matter ($PM_{2.5}$); concentration-weighted trajectory (CWT); XGBoost model; regional transport

1. Introduction

Ambient fine particulate matter ($PM_{2.5}$, with particle aerodynamic diameter $\leq$ 2.5 μm) is influenced by both natural sources, such as dust, volcanic eruptions, tsunamis, and forest fires, and anthropogenic emissions, including fuel combustion, transportation, and industrial production. Anthropogenic emissions dominate the long-term trend of air pollution (Zhang et al., 2019; Cheng et al., 2019). Numerous epidemiological studies have found that $PM_{2.5}$ can significantly damage human health by exacerbating respiratory and cardiovascular diseases (Bartell et al., 2013; Brauer et al., 2012; Pascal et al., 2014), and also has an impact on weather and climate change (Wang et al., 2014; Smith et al., 2020; Kalisoras et al., 2023). China's rapid and energy-intensive development over the past several decades has led to severe air pollution and negative public health impacts (Huang et al., 2014; Geng et al., 2021). Consequently, controlling pollution and reducing $PM_{2.5}$ concentrations became an urgent issue in China. While meteorological variations caused about 16% of the ambient $PM_{2.5}$ decline during 2013-2017 (Zhang et al., 2019), the uncertainty in reducing $PM_{2.5}$ through meteorological conditions is substantial, and the magnitude of the decrease is not dominated by human actions. Thus, the primary means of controlling $PM_{2.5}$ relies on reducing anthropogenic emissions. To address this issue, the Chinese government implemented the Air Pollution Prevention and Control Action Plan (denoted "Action Plan") from 2013 to 2017 and the Blue Sky Protection Campaign from 2018 to 2020, which effectively controlled anthropogenic emissions and reduced ambient $PM_{2.5}$ concentrations (Zhang et al., 2019; Du et al., 2022).

The concentration of $PM_{2.5}$ can be attributed to local emissions and regional transport. Several methods, such as the HYSPLIT model (Draxler and Rolph, 2010), can be used to distinguish pollutant sources. Wu et al. used the HYSPLIT model to simulate the 24-hour backward trajectory in Zhoushan (Wu et al., 2021), and identified continental air masses that spent more than 5% of the previous 24 hours over the continent region, while the remaining air masses were identified as oceanic-influenced air masses. Ding et al. employed a backward trajectory ensemble to analyze the sources of air masses in Beijing during the study period (Ding et al., 2019), finding that air masses with high concentrations of black carbon (BC) mass mainly came from the south and southeast regions. Cluster analysis on backward trajectories can be used to obtain the main direction of aerosols over a period of time, allowing for the analysis and determination of dominant air mass directions. For instance, Li et al. divided the sources of air masses in the Wuhan area from October to November 2019 into short transport distance, northbound air masses, and regional transport from the northeast and some coastal areas (Li et al., 2022).

The HYSPLIT model results are mainly used to view air mass trajectories, making it difficult to directly determine the sources of pollutants. Potential source contribution function (PSCF) and concentration-weighted trajectory (CWT) analyses based on backward trajectories can be used to identify the sources of pollutants through conditional probability results. Hu et al. used weighted PSCF to analyze the sources of air masses with different levels of pollution in Beijing and found that polluted air masses from the southwest were an important source of high-level advections during the study period, while light pollution was often accompanied by the regional transport originating from the northeast region (Hu et al., 2020). Wu et al. used CWT to analyze the sources of pollution in Zhoushan and found that pollutants in Zhoushan are influenced by both local emissions and regional transport. There are no obvious high pollution areas, while in other seasons, $PM_{2.5}$ mainly originates from southern Jiangsu and Shanghai (Wu et al., 2024). However, these studies relied on standard HYSPLIT trajectory results, which have lower temporal resolution, limiting the accuracy of pollutant source identification.

The Lagrangian air pollution dispersion model, Numerical Atmospheric-dispersion Modelling Environment (NAME) (Jones et al., 2007) can determine the source of polluted air masses by simulating particulate concentrations within each grid point using Monte Carlo methods, followed by 3-D trajectories of plume basins. Liu et al. used the NAME model to study the sources of air masses in Beijing during the winter of 2019 and divided them into local emissions and regional transport to analyze the convective mixing process of BC under the influence of local emissions (Liu et al., 2020). However, due to limitations in computing resources, the NAME model is difficult to use for obtaining long-term emission source analysis results.

Multiple methods can be used to predict $PM_{2.5}$ concentrations, such as statistical models (e.g., linear mixed-effect models and generalized additive models) (Fang et al., 2016; Ma et al., 2016), chemical transport model (CTM)-based algorithms (Geng et al., 2015; Kong et al., 2021), physical models (Lin et al., 2018), and recently emerging machine learning models, including Extreme Gradient Boosting (XGBoost) and Random Forest (Liang et al., 2020; Wei et al., 2021; Xiao et al., 2018; Xue et al., 2019; Huang et al., 2021). Geng et al. used satellite observations of aerosol optical depth (AOD) and meteorological data combined with the XGBoost model to explore the long-term variations of $PM_{2.5}$ caused by changes in meteorological conditions from 2000 to 2018 (Geng et al., 2021). Kleine Deters et al. demonstrated the relevance of statistical models based on machine learning for predicting $PM_{2.5}$ concentrations from meteorological data (Kleine Deters et al., 2017). This method of predicting aerosol concentrations using only meteorological data has been widely used (Asadollahfardi et al., 2016; Zeng et al., 2021). For instance, Grange et al. used meteorological data, synoptic scale weather patterns, and time variables to explain daily PM_{10} concentrations in Switzerland (Grange et al., 2018). In summary, machine learning models have achieved high accuracy in estimating and predicting $PM_{2.5}$ concentrations and have high use value, and the rise of machine learning methods has also provided feasibility for quantifying the contribution of regionally transported air masses.

In this study, we combined CWT analysis with the HYSPLIT trajectory ensemble to obtain hourly-resolution $PM_{2.5}$ source results and used this approach to distinguish between local emissions and regional transport. Solved the problems of traditional CWT methods being unable to obtain hourly time accuracy and models such as NAME consuming a large number of computational resources. Predictive XGBoost models were developed for Beijing using meteorological data and time variables to explain $PM_{2.5}$ concentrations. By training the XGBoost model with $PM_{2.5}$ dominated by local emissions, which are separately distinguished by CWT, and generalizing the findings to all study periods, the concentration of locally emitted $PM_{2.5}$ (local) can be obtained. Similarly, ambient observed $PM_{2.5}$ (ambient) can be determined by training the XGBoost model with ambient $PM_{2.5}$ data. The contribution of regional transport to $PM_{2.5}$ in Beijing can be quantified by comparing the ambient and local $PM_{2.5}$ concentrations.

2. Materials and methods

2.1 Site and instrumentation

The $PM_{2.5}$ data (Fig. 1a) were obtained from in situ air quality monitoring conducted by the China National Environmental Monitoring Center from 2013 to 2020. The monitoring station is located in Haidian Wanliu (39.96°N, 116.29°E), situated in the central urban area of Beijing. Meteorological data, including temperature, relative humidity, pressure, precipitation, wind speed, and planetary boundary layer height (PBLH), were sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 hourly reanalysis dataset (https://cds.climate.copernicus.eu/datasets). In this study, a year was divided into four quarters: Spring (March, April, and May), Summer (June, July, and August), Autumn (September, October, and November), and Winter (December, January, and February).

2.2 Air mass source

The air mass trajectory data were obtained from the 1°×1° horizontal and vertical wind fields of the Global Data Assimilation System (GDAS) reanalysis products (ftp://arlftp.arlhq.noaa.gov/pub/archives/gdas1), which are available every 3 hours. The HYSPLIT trajectory ensemble was used to generate 27 equally probable 24-hour backward air mass trajectories for the target point (39.96°N, 116.29°E, 250 m a.s.l.) in every hour by using PySplit (Cross, 2015). Given the equal probability of air masses being transported to the target point for each trajectory in the HYSPLIT trajectory ensemble, a conditional probability CWT analysis was applied to determine the hourly source area of pollution.

In the CWT analysis method, each grid point is assigned a weight, and the contribution of each grid point to the pollutant concentration at the target site is calculated using the air mass residence time and pollutant concentration (Hopke et al., 1993; Polissar et al., 1999; Xu and Akhtar, 2010) (equation 1). The grid point resolution was set to 0.25°×0.25° for this study. In equations 1, C_{ij} is the average weighted concentration at grid point (*i, j*), *l* is the trajectory index, M represents the total number of trajectories, C_l is the $PM_{2.5}$ concentration corresponding to the target site, and τ_{ijl} is the residence time of trajectory *l* passing through the grid point. In calculation, the number of trajectories falling on each grid point is used instead of the residence time.

$$C_{ij} = \frac{\sum_{l=1}^{M} C_l \times \tau_{ijl}}{\sum_{l=1}^{M} \tau_{ijl}} \tag{1}$$

To reduce the effect of small values of n_{ij}, the CWT values were multiplied by an arbitrary weight function $W(n_{i,j})$ to better reflect the uncertainty in the values for these grids (equation 2).

$$W(n_{i,j}) = \begin{cases} 1.00,\ 3n_{ave} < n_{ij} \\ 0.70, 1.5n_{ave} < n_{ij} \le 3n_{ave} \\ 0.4, n_{ave} < n_{ij} \le 1.5n_{ave} \\ 0.17, n_{ij} \le n_{ave} \end{cases} \tag{2}$$

where n_{ij} represents the number of trajectories that fall within the grid point, and n_{ave} represents the average number of trajectories passing through each grid point.

The potential source contribution to $PM_{2.5}$ at the target site was investigated by categorizing the backward air masses into five different source regions centered around Beijing: local (which is a region around central Beijing, 115.3~117.5°E, 39.4~41°N); north region (the northern plateau at 108~117.5°E, 41~43°N); west region (the western plateau at 108~115.3°E, 34~41°N); south region (the southern plain at 115.3~120°E, 34~39.4°N); and east region (the eastern plain at 117.5~120°E, 39.4~43°N). The concentration is integrated over each grid point in each segregated region obtained from the CWT analysis, and the contributions of each air mass fraction are obtained. The region with the highest contribution is used to determine the dominant source of air masses in Beijing at each time, classifying the overall air mass sources into local emissions (Fig. 1g) and regional transport (Fig. 1h). It is important to note that local emission periods were also influenced by persistent regional transport, and vice versa.

2.3 Deriving the long-term local emission and ambient $PM_{2.5}$

An XGBoost model is employed to derive the local and ambient $PM_{2.5}$ results. The hyperparameters used in the model for local (ambient) conditions include a maximum number of boosting iterations of 6067 (13421), a learning rate of 0.1, a maximum tree depth of 7 (11), a minimum sum of instance weight needed in a child of 5 (3), a subsampling ratio of 0.8 (0.6) for training instances, and a subsampling ratio of 0.8 for columns when constructing each tree. The input parameters for the XGBoost model comprise meteorological variables (temperature, relative humidity, wind speed, surface pressure, and precipitation) and temporal parameters (year, month, day of the week, and day of the year), as referenced from Xu et al. (Xu et al., 2023). Additionally, PBLH, which has been shown to significantly impact pollutant concentrations in previous observational (Su et al., 2018; Miao and Liu, 2019; Miao et al., 2019) and machine learning studies (Xiao et al., 2021; Li et al., 2017b; Shen et al., 2018), was included as an input parameter. Based on the XGBoost learning results, the most sensitive parameters for both local and ambient $PM_{2.5}$ are RH, wind field, surface pressure and PBLH (Fig. S1). For the machine learning process, data from 2013 to 2019 were used for training the XGBoost models, while data from 2020 were used for model validation. Note that the 2020 analysis results may contain some uncertainties due to the impact of COVID-19.

The relatively small proportion of high-concentration $PM_{2.5}$ can lead to underestimation of high-concentration events in the model results (Wei et al., 2020). To address this issue, a high $PM_{2.5}$ indicator was defined as a daily average $PM_{2.5}$ concentration exceeding the monthly average plus twice the standard deviation. In this study, original high $PM_{2.5}$ indicators accounted for 6% of the data points during the period dominated by local and ambient $PM_{2.5}$. To balance the proportion of high-concentration $PM_{2.5}$ in the entire database, the Synthetic Minority Over-sampling Technique (SMOTE) (Torgo, 2011) was applied during data preprocessing. SMOTE artificially generates new synthetic samples along the line between high-concentration data points and their selected nearest neighbors, effectively oversampling the high-concentration data. As a result, the proportion of high $PM_{2.5}$ indicators increased to 21% and 22% for local and ambient $PM_{2.5}$, respectively.

Hyperparameter optimization and performance evaluation of the model were conducted using fivefold cross-validation (CV), while early stopping with a patience of 10 rounds was employed to prevent overfitting. (Akritidis et al., 2021; Zhang et al., 2020). In this approach, 20% of the data is randomly selected for model testing, while the remaining 80% is used for training. This process is repeated five times, ensuring that each record is used once as testing data. The coefficient of determination (r^2) was employed to assess the correlation between the XGBoost model predictions and observed values, while the root mean square error (RMSE) was used as a performance evaluation statistic. After obtaining the relation between the input parameters and $PM_{2.5}$, we are able to derive the hourly local and ambient $PM_{2.5}$ once all long-term input parameters (Fig. S3).

3 Results and discussion

3.1 Evaluation of the XGBoost $PM_{2.5}$ prediction model

During the model validation process, the XGBoost model results for ambient $PM_{2.5}$ (Fig. 2a2) demonstrated an r^2 of 0.74 and an RMSE of 20 $\mu g\ m^{-3}$ when compared to observations. The XGBoost model results for local $PM_{2.5}$ exhibited an r^2 of 0.78 and an RMSE of 21 $\mu g\ m^{-3}$. An analysis of the $PM_{2.5}$ frequency distribution in Beijing revealed an agreement between the XGBoost model results and observations for both ambient and local $PM_{2.5}$. The error between XGBoost learning results and actual observed $PM_{2.5}$ values is mainly concentrated in the low concentration stage. This may be attributed to the significant reduction in human activities during the COVID-19 lockdown periods, which led to a decrease in actual $PM_{2.5}$ levels, making it challenging for XGBoost to learn (Fig. 2b1 and b2). As illustrated in Fig. S2, local and ambient $PM_{2.5}$ in Beijing display a distinct seasonal variation, with higher values in winter and lower values in summer. However, the transport of clean air masses from the north diminishes the seasonal variation characteristics of ambient $PM_{2.5}$ in Beijing, making winter pollution less prominent compared to other seasons.

Fig. S3 reveals that ambient pollution events ($PM_{2.5}$ > 75 $\mu g\ m^{-3}$) in Beijing are primarily influenced by air masses originating from the south and west, particularly under the control of westward air masses. Numerous studies have indicated that air masses originating from the western region significantly contribute to regional pollution events in Beijing (Streets et al., 2007; Tian et al., 2019; Liu et al., 2020). With the exception of December (Fig. 3b1), westward air masses often bring higher monthly average $PM_{2.5}$ to Beijing. Air masses originating from the south region can also transport more pollutants to Beijing (Fig. S3). However, unlike the high-frequency polluted air masses from the west, southward air masses are associated with higher $PM_{2.5}$ concentrations, particularly during autumn and winter (Fig. 3c1). This phenomenon can be attributed to the higher pollution levels in Hebei and Shandong provinces compared to Beijing during these seasons, as verified by AOD observations from Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua satellites over Eastern China (Zhang and Reid, 2010; Hu et al., 2018) (Fig. S4). Notably, in contrast to westward transport, air masses from the south region in February predominantly exhibited a cleaning effect on Beijing, even before 2017 (Fig. S3b). This can be explained by the occurrence of these transport processes during or shortly after the Spring Festival, a period characterized by extremely low anthropogenic emissions, resulting in lower ambient $PM_{2.5}$ compared to local emissions in the megacity of Beijing. Following the implementation of the Action Plan, the polluted air masses from the south region transitioned from carrying higher $PM_{2.5}$ to levels close to local emission concentrations in Beijing, leading to a more equal contribution to pollution and clean events in the area (Fig. S5c1).

3.2 Impact of clean air masses from transported regions on $PM_{2.5}$ in Beijing

In this study, clean air masses are defined as those associated with ambient $PM_{2.5}$ in the Beijing area that are lower than the concentrations resulting from local emissions, as illustrated below the dashed line in Fig. 3a1-d1. This study reveals that clean air masses predominantly originate from the east and north regions during the period 2013-2020, which is consistent with previous studies (Zhang et al., 2018; Hu et al., 2020). Clean air masses from different directions exhibit similar seasonal variations in their ability to reduce locally emitted pollution in Beijing, with a strong reduction effect in winter and a weaker effect in summer (Fig. 3a2-d2). This phenomenon is closely related to the seasonal variations in pollutant emissions. Due to the combined influence of increased residential emissions from heating activities and meteorological conditions in Beijing during autumn and winter, local $PM_{2.5}$ in Beijing presents higher concentrations. Consequently, the influx of clean air masses results in a more pronounced reduction in $PM_{2.5}$ during these seasons. The weaker attenuation effect of $PM_{2.5}$ transported from the south region during December and January can be attributed to the high-frequency and high-concentration pollution contributions from air masses originating in this region during this period.

Due to a significant reduction in anthropogenic emissions after 2017, the attenuation of $PM_{2.5}$ concentrations by clean air masses from all directions was significantly lower than before 2017 (Fig. S6a2-d2). Compared to the period prior to 2017, the mean attenuation of $PM_{2.5}$ concentrations in Beijing decreased by 3, 10, 3, and 7 μg m^{-3} ($p < 0.01$) for air masses originating from the north, west, south, and east regions, respectively.

3.3 Variations in Beijing $PM_{2.5}$ concentrations under transport-induced pollution events

Transport-induced pollution events in Beijing are defined as the occurrence of ambient $PM_{2.5}$ exceeding both local $PM_{2.5}$ and the light pollution standard (75 μg m^{-3}). Fig. 4a1-d1 demonstrate that the monthly variation of $PM_{2.5}$ in Beijing generally follows a unimodal pattern, with higher values in winter and lower values in summer, except when under the influence of eastern air mass transport. This phenomenon is closely related to the seasonal variations in anthropogenic emissions in China and the characteristics of climate change (Renhe et al., 2014; Li et al., 2017a; Zhang et al., 2015). The overall $PM_{2.5}$ in Beijing under the influence of eastward pollution air masses exhibits a bimodal distribution, with frequent high-concentration pollution events occurring in January and October. Even after the effective control of anthropogenic emissions in 2017, a second peak of high-concentration pollution persists in October (Fig. 4d2). Fig. 4a2-d2 illustrate the effectiveness of the Action Plan in controlling pollutant concentrations in the Beijing area. Since 2017, $PM_{2.5}$ in Beijing has been significantly lower than the values observed before 2017 during transport-induced pollution events. Moreover, during January and from

June to September, there were periods when the regional transport of polluted air masses from a fixed direction did not contribute to pollution events in Beijing.

An analysis of the proportion of transport-induced pollution events from different regions to Beijing (Fig. 5) shows that after the implementation of the Action Plan in 2017, the number of pollution events dominated by regional transport decreased significantly. From spring to winter, the largest decrease in transport-induced pollution events occurred in the north, west, west and south regions in each season, with the lowest decrease occurring in the east region during winter.

The temporal variation in the number of transport-induced pollution events from different regions (Fig. S7) revealed that air masses transported from the west region contributed to the most frequent pollution events in each season except summer. The highest number of events occurred in spring 2016 (322), autumn 2016 (375), and winter 2017 (308). Summer transport-induced pollution events were mainly influenced by polluted air masses transported from the south, with a gradual decrease in the number of events over the years. Although pollution events in Beijing primarily occur in autumn and winter, this study found that after 2017, the season when Beijing was most affected by transport-induced pollution events was spring, contributing a total of 685 pollution events, while autumn and winter contributed 266 and 392 events, respectively. The impact of polluted air masses on summer transport was minimal, with only 215 occurrences.

Fig. 5a shows that in spring, transport-induced pollution events in Beijing were mainly dominated by polluted air masses transported from the west and south. The highest proportion of regional transport events from the west occurred in 2016, reaching 68%, while the highest proportion of southward transport-induced pollution events occurred in 2017 (with the exception of 2020, which may have been influenced by the COVID-19 pandemic). The increased frequency of pollution air masses transported from the south after 2017 can be attributed to the effective control of anthropogenic emissions, resulting in a decrease in $PM_{2.5}$ transported from various regions, especially from westward sources (Fig. S7a). The decrease in the proportion of pollution events transported from the west, which originally accounted for a large proportion, led to an increase in the contribution of remaining incoming air masses to Beijing.

Before 2017, transport-induced pollution events in Beijing during summer were mainly affected by polluted air masses from the south region. Even in 2015, when the proportion of transport-induced pollution events from south region was lowest during the entire period, it still accounted for 50% of the total number of transport-induced pollution events that year. However, after the implementation of the Action Plan, the proportion of transport-induced pollution events from the south region gradually decreased to 38%. In 2020, this proportion further declined to 25%, but this may have been affected by the COVID-19 pandemic. Meanwhile, pollution air masses originating from the east increasingly dominated the occurrence of pollution events in Beijing.

Transport-induced pollution events in Beijing mainly originated from the west and had the highest contribution proportion in autumn before 2019 (except for 2013, when the contribution proportion was 34%, second only to southward air masses at 35%). After 2019, the contribution of eastward air masses became dominant in autumn. In winter, polluted air masses from the west were the main source of transport-induced pollution events. Overall, as the Action Plan gradually improved, the transport-induced pollution from the east did not decrease significantly compared to other air mass sources. This may be because the eastward air masses are mostly clean. However, as the concentration of polluted air masses from other sources decreases, the potential impact of eastward air masses on Beijing's transport-induced pollution events increases. This finding may prompt Beijing to prioritize emission reduction in the east region when implementing future joint prevention and control measures.

4 Conclusion

This study combined a machine learning method and Concentration-Weighted Trajectory (CWT) analysis to derive local emissions and ambient observed $PM_{2.5}$ in Beijing from 2013 to 2020, thus the contribution of regional transport to $PM_{2.5}$ in Beijing can be quantified. The impact of clean air masses (defined as those with ambient $PM_{2.5}$ concentrations lower than local emissions) mainly originated from the east and north regions. These clean air masses from different directions exhibited similar seasonal variations in their ability to reduce ambient pollution in Beijing, with a stronger reduction effect in winter and a weaker reduction effect in summer.

Except for the regional transport from the east region, the seasonal variation of $PM_{2.5}$ in Beijing under the influence of transport-induced pollution events (ambient $PM_{2.5}$ exceeding both local $PM_{2.5}$ and 75 μg m^{-3}) shows a general trend of high concentrations in winter and low concentrations in summer. The main reason for this phenomenon is related to the seasonal emissions of pollutants in China and the characteristics of climate change. Before 2019, the west region was the primary source of pollution events during autumn and winter. However, starting from 2019, the east region became the main contributor of polluted air masses in autumn. Additionally, among all regions, the east region exhibited the smallest decrease in transport-induced pollution events after 2017.

From 2013 to 2020, local emissions were the main contributors to pollution events in Beijing. However, the Air Pollution Prevention and Control Action Plan, implemented by the Chinese government in 2017, more effectively mitigated pollutants caused by regional transport compared to local emissions, particularly during autumn and winter. This finding suggests that Beijing should prioritize reducing local emissions while also accounting for potential contributions from the east region in its future pollution prevention and control strategies.

Code and data availability

The codes used in this study are archived on Zenodo: the machine learning code at https://doi.org/10.5281/zenodo.14677125, the CWT code at https://doi.org/10.5281/zenodo.13994400, ECMWF data at https://doi.org/10.5281/zenodo.14353871, GDAS data at https://doi.org/10.5281/zenodo.14347277, HySplit Trajectory Ensemble at https://doi.org/10.5281/zenodo.14375567, and PySPLIT at https://doi.org/10.5281/zenodo.14354765. The meteorology and $PM_{2.5}$ data used in this study can be accessed at https://dx.doi.org/10.17632/bhfktx3kz8.2.

Author contribution

Kang Hu, Hong Liao and Dantong Liu designed and carried out the experiments. Kang Hu wrote the code and final paper with contributions from all other authors. Hong Liao, Dantong Liu, Lei Chen and Jianbing Jin reviewed and edited the paper.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This research was supported by the China Postdoctoral Science Foundation (2023M741773), Postdoctoral Fellowship Program of CPSF (GZC20231150), National Natural Science Foundation of China (42021004, 42405192).

Reference

Akritidis, D., Zanis, P., Georgoulias, A. K., Papakosta, E., Tzoumaka, P., and Kelessis, A.: Implications of COVID-19 restriction measures in urban air quality of Thessaloniki, Greece: A machine learning approach, Atmosphere, 12, 1500, 2021.

Asadollahfardi, G., Madinejad, M., Aria, S. H., and Motamadi, V.: Predicting Particulate Matter (PM2. 5) Concentrations in the Air of Shahr-e Ray City, Iran, by Using an Artificial Neural Network, Environmental Quality Management, 25, 71-83, 2016.

Bartell, S. M., Longhurst, J., Tjoa, T., Sioutas, C., and Delfino, R. J.: Particulate air pollution, ambulatory heart rate variability, and cardiac arrhythmia in retirement community residents with coronary artery disease, Environmental health perspectives, 121, 1135-1141, 2013.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., and Van Dingenen, R.: Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution, Environmental science & technology, 46, 652-660, 2012.

Cheng, N., Cheng, B., Li, S., and Ning, T.: Effects of meteorology and emission reduction measures on air pollution in Beijing during heating seasons, Atmospheric Pollution Research, 10, 971-979, 2019.

Cross, M.: PySPLIT: a Package for the Generation, Analysis, and Visualization of HYSPLIT Air Parcel Trajectories, SciPy, 133-137,

Ding, S., Zhao, D., He, C., Huang, M., He, H., Tian, P., Liu, Q., Bi, K., Yu, C., and Pitt, J.: Observed interactions between black carbon and hydrometeor during wet scavenging in mixed-phase clouds, Geophysical Research Letters, 46, 8453-8463, 2019.

Draxler, R. and Rolph, G.: HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) model access via NOAA ARL READY website (http://ready. arl. noaa. gov/HYSPLIT. php). NOAA Air Resources Laboratory, Silver Spring, MD, 25, 2010.

Du, H., Li, J., Wang, Z., Chen, X., Yang, W., Sun, Y., Xin, J., Pan, X., Wang, W., and Ye, Q.: Assessment of the effect of meteorological and emission variations on winter PM2. 5 over the North China Plain in the three-year action plan against air pollution in 2018–2020, Atmospheric Research, 280, 106395, 2022.

Fang, X., Zou, B., Liu, X., Sternberg, T., and Zhai, L.: Satellite-based ground PM2. 5 estimation using timely structure adaptive modeling, Remote Sensing of Environment, 186, 152-163, 2016.

Geng, G., Zhang, Q., Martin, R. V., van Donkelaar, A., Huo, H., Che, H., Lin, J., and He, K.: Estimating long-term PM2. 5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model, Remote sensing of Environment, 166, 262-270, 2015.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., and Peng, Y.: Tracking air pollution in China: near real-time PM2. 5 retrievals from multisource data fusion, Environmental Science & Technology, 55, 12106-12115, 2021.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM 10 trend analysis, Atmospheric Chemistry and Physics, 18, 6223-6239, 2018.

Hopke, P. K., Gao, N., and Cheng, M.-D.: Combining chemical and meteorological data to infer source areas of airborne pollutants, Chemometrics and Intelligent Laboratory Systems, 19, 187-199, 1993.

Hu, K., Kumar, K. R., Kang, N., Boiyo, R., and Wu, J.: Spatiotemporal characteristics of aerosols and their trends over mainland China with the recent Collection 6 MODIS and OMI satellite datasets, Environmental Science and Pollution Research, 25, 6909-6927, 2018.

Hu, K., Zhao, D., Liu, D., Ding, S., Tian, P., Yu, C., Zhou, W., Huang, M., and Ding, D.: Estimating radiative impacts of black carbon associated with mixing state in the lower atmosphere over the northern North China Plain, Chemosphere, 252, 126455, 2020.

Huang, C., Hu, J., Xue, T., Xu, H., and Wang, M.: High-resolution spatiotemporal modeling for ambient PM2. 5 exposure assessment in China from 2013 to 2019, Environmental Science & Technology, 55, 2152-2162, 2021.

Huang, R.-J., Zhang, Y., Bozzetti, C., Ho, K.-F., Cao, J.-J., Han, Y., Daellenbach, K. R., Slowik, J. G., Platt, S. M., and Canonaco, F.: High secondary aerosol contribution to particulate pollution during haze events in China, Nature, 514, 218-222, 2014.

Jones, A., Thomson, D., Hort, M., and Devenish, B.: The UK Met Office's next-generation atmospheric dispersion model, NAME III, in: Air pollution modeling and its application XVII, Springer, 580-589, 2007.

Kalisoras, A., Georgoulias, A. K., Akritidis, D., Allen, R. J., Naik, V., Kuo, C., Szopa, S., Nabat, P., Olivié, D., and Van Noije, T.: Decomposing the Effective Radiative Forcing of anthropogenic aerosols based on CMIP6 Earth System Models, Atmospheric Chemistry & Physics Discussions, 2023.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., and Rybarczyk, Y.: Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters, Journal of Electrical and Computer Engineering, 2017, 5106045, 2017.

Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., Wu, Q., Chen, H., Zhu, L., and Wang, W.: A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC, Earth System Science Data, 13, 529-570, 2021.

Li, M., Zhang, Q., Kurokawa, J.-i., Woo, J.-H., He, K., Lu, Z., Ohara, T., Song, Y., Streets, D. G., and Carmichael, G. R.: MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP, Atmospheric Chemistry and Physics, 17, 935-963, 2017a.

Li, S., Liu, D., Kong, S., Wu, Y., Hu, K., Zheng, H., Cheng, Y., Zheng, S., Jiang, X., and Ding, S.: Evolution of source attributed organic aerosols and gases in a megacity of central China, Atmospheric Chemistry and Physics Discussions, 2022, 1-19, 2022.

Li, T., Shen, H., Yuan, Q., Zhang, X., and Zhang, L.: Estimating ground-level PM2. 5 by fusing satellite and station observations: a geo-intelligent deep learning approach, Geophysical Research Letters, 44, 11,985-911,993, 2017b.

Liang, F., Xiao, Q., Huang, K., Yang, X., Liu, F., Li, J., Lu, X., Liu, Y., and Gu, D.: The 17-y spatiotemporal trend of PM2. 5 and its mortality burden in China, Proceedings of the National Academy of Sciences, 117, 25601-25608, 2020.

Lin, C., Liu, G., Lau, A. K. H., Li, Y., Li, C., Fung, J. C. H., and Lao, X. Q.: High-resolution satellite remote sensing of provincial PM2. 5 trends in China from 2001 to 2015, Atmospheric environment, 180, 110-116, 2018.

Liu, D., Hu, K., Zhao, D., Ding, S., Wu, Y., Zhou, C., Yu, C., Tian, P., Liu, Q., and Bi, K.: Efficient vertical transport of black carbon in the planetary boundary layer, Geophysical Research Letters, 47, e2020GL088858, 2020.

Ma, Z., Hu, X., Sayer, A. M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., and Liu, Y.: Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004–2013, Environmental health perspectives, 124, 184-192, 2016.

Miao, Y. and Liu, S.: Linkages between aerosol pollution and planetary boundary layer structure in China, Science of the Total Environment, 650, 288-296, 2019.

Miao, Y., Li, J., Miao, S., Che, H., Wang, Y., Zhang, X., Zhu, R., and Liu, S.: Interaction between planetary boundary layer and PM 2.5 pollution in megacities in China: a Review, Current Pollution Reports, 5, 261-271, 2019.

Pascal, M., Falq, G., Wagner, V., Chatignoux, E., Corso, M., Blanchard, M., Host, S., Pascal, L., and Larrieu, S.: Short-term impacts of particulate matter (PM10, PM10–2.5, PM2. 5) on mortality in nine French cities, Atmospheric Environment, 95, 175-184, 2014.

Polissar, A., Hopke, P., Paatero, P., Kaufmann, Y., Hall, D., Bodhaine, B., Dutton, E., and Harris, J.: The aerosol at Barrow, Alaska: long-term trends and source locations, Atmospheric Environment, 33, 2441-2458, 1999.

Renhe, Z., Li, Q., and Zhang, R.: Meteorological conditions for the persistent severe fog and haze event over eastern China in January 2013, Science China Earth Sciences, 57, 26-35, 2014.

Shen, H., Li, T., Yuan, Q., and Zhang, L.: Estimating regional ground-level PM2. 5 directly from satellite top-of-atmosphere reflectance using deep belief networks, Journal of Geophysical Research: Atmospheres, 123, 13,875-813,886, 2018.

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., Boucher, O., Dufresne, J.-L., Nabat, P., and Michou, M.: Effective radiative forcing and adjustments in CMIP6 models, Atmospheric Chemistry and Physics, 20, 9591-9618, 2020.

Streets, D. G., Fu, J. S., Jang, C. J., Hao, J., He, K., Tang, X., Zhang, Y., Wang, Z., Li, Z., and Zhang, Q.: Air quality during the 2008 Beijing Olympic games, Atmospheric environment, 41, 480-492, 2007.

Su, T., Li, Z., and Kahn, R.: Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: regional pattern and influencing factors, Atmospheric Chemistry and Physics, 18, 15921-15935, 2018.

Tian, P., Liu, D., Huang, M., Liu, Q., Zhao, D., Ran, L., Deng, Z., Wu, Y., Fu, S., and Bi, K.: The evolution of an aerosol event observed from aircraft in Beijing: An insight into regional pollution transport, Atmospheric Environment, 206, 11-20, 2019.

Torgo, L.: Data mining with R: learning with case studies, chapman and hall/CRC2011.

Wang, Y., Wang, M., Zhang, R., Ghan, S. J., Lin, Y., Hu, J., Pan, B., Levy, M., Jiang, J. H., and Molina, M. J.: Assessing the effects of anthropogenic aerosols on Pacific storm track using a multiscale global climate model, Proceedings of the National Academy of Sciences, 111, 6894-6899, 2014.

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

Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., and Song, Y.: Improved 1 km resolution PM2.5 estimates across China using enhanced space–time extremely randomized trees, Atmos. Chem. Phys., 20, 3273-3289, 10.5194/acp-20-3273-2020, 2020.

Wu, Y., Liu, D., Wang, X., Li, S., Zhang, J., Qiu, H., Ding, S., Hu, K., Li, W., and Tian, P.: Ambient marine shipping emissions determined by vessel operation mode along the East China Sea, Science of The Total Environment, 769, 144713, 2021.

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data, Environmental science & technology, 52, 13260-13269, 2018.

Xiao, Q., Zheng, Y., Geng, G., Chen, C., Huang, X., Che, H., Zhang, X., He, K., and Zhang, Q.: Separating emission and meteorological contributions to long-term PM_{2.5} trends over eastern China during 2000–2018, Atmospheric Chemistry and Physics, 21, 9475-9496, 10.5194/acp-21-9475-2021, 2021.

Xu, R., Ye, T., Yue, X., Yang, Z., Yu, W., Zhang, Y., Bell, M. L., Morawska, L., Yu, P., and Zhang, Y.: Global population exposure to landscape fire air pollution from 2000 to 2019, Nature, 621, 521-529, 2023.

Xu, X. and Akhtar, U.: Identification of potential regional sources of atmospheric total gaseous mercury in Windsor, Ontario, Canada using hybrid receptor modeling, Atmospheric Chemistry and Physics, 10, 7073-7083, 2010.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., and Zhang, Q.: Spatiotemporal continuous estimates of PM2. 5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations, Environment international, 123, 345-357, 2019.

Zeng, Z., Gui, K., Wang, Z., Luo, M., Geng, H., Ge, E., An, J., Song, X., Ning, G., and Zhai, S.: Estimating hourly surface PM2. 5 concentrations across China from high-density meteorological observations by machine learning, Atmospheric Research, 254, 105516, 2021.

Zhang, J. and Reid, J.: A decadal regional and global trend analysis of the aerosol optical depth using a data-assimilation grade over-water MODIS and Level 2 MISR aerosol products, Atmospheric Chemistry and Physics, 10, 10949-10963, 2010.

Zhang, L., Wang, T., Lv, M., and Zhang, Q.: On the severe haze in Beijing during January 2013: Unraveling the effects of meteorological anomalies with WRF-Chem, Atmospheric Environment, 104, 11-21, 2015.

Zhang, L., Zhao, T., Gong, S., Kong, S., Tang, L., Liu, D., Wang, Y., Jin, L., Shan, Y., and Tan, C.: Updated emission inventories of power plants in simulating air quality during haze periods over East China, Atmospheric Chemistry and Physics, 18, 2065-2079, 2018.

Zhang, Q., Wu, S., Wang, X., Sun, B., and Liu, H.: A PM2. 5 concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations, Journal of Cleaner Production, 275, 122722, 2020.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., Xu, X., Wang, J., He, H., and Liu, W.: Drivers of improved PM2. 5 air quality in China from 2013 to 2017, Proceedings of the National Academy of Sciences, 116, 24463-24469, 2019.

Figures and captions

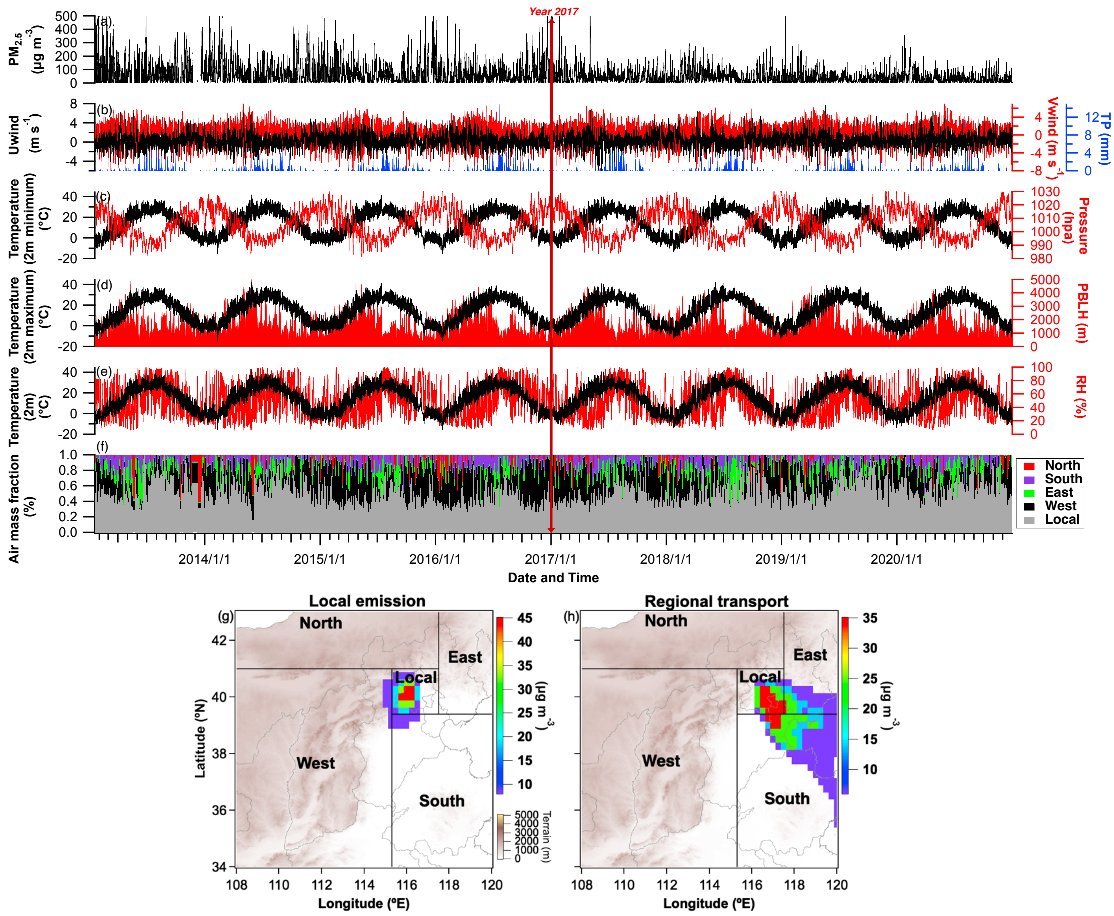

Fig. 1. Temporal evolution of parameters used in the XGBoost model: (a) $PM_{2.5}$; (b) U-wind, V-wind, and total precipitation; (c) 2-m minimum temperature and surface pressure; (d) 2-m maximum temperature and planetary boundary layer height; (e) 2-m temperature and relative humidity; (f) air mass fraction in contributing sources derived from the Concentration-Weighted Trajectory (CWT) model for a 1-day backward trajectory. The red vertical line with arrows indicates the implementation of environmental regulations. Typical examples of the CWT model analysis are shown for (g) a local emission period (25 August 2013) and (h) a regional transport period (15 July 2013).

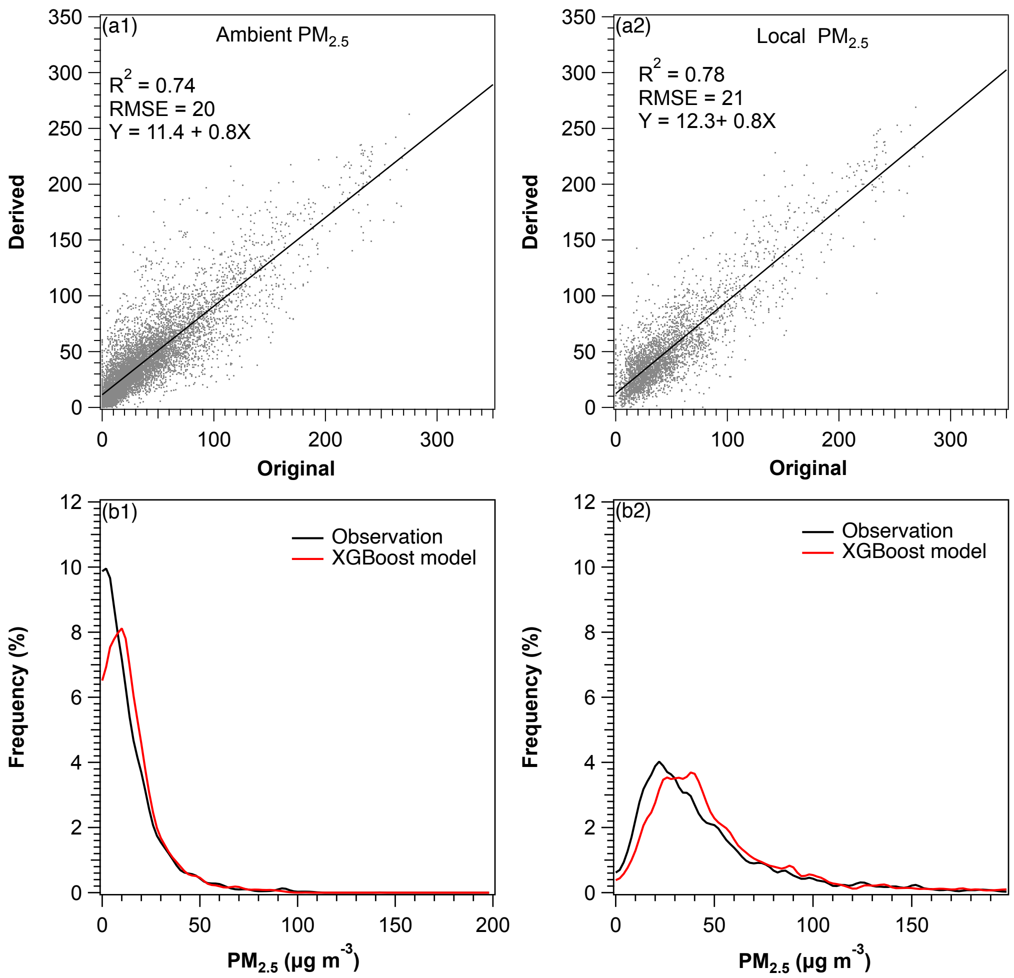

Fig. 2. Comparison of XGBoost model estimates and observations for (a1) ambient $PM_{2.5}$ and (a2) local $PM_{2.5}$ using testing samples from 2020. Frequency distributions of $PM_{2.5}$ observations (black lines) and XGBoost model predictions (red lines) for (b1) ambient $PM_{2.5}$ and (b2) local $PM_{2.5}$ using testing samples from 2020.

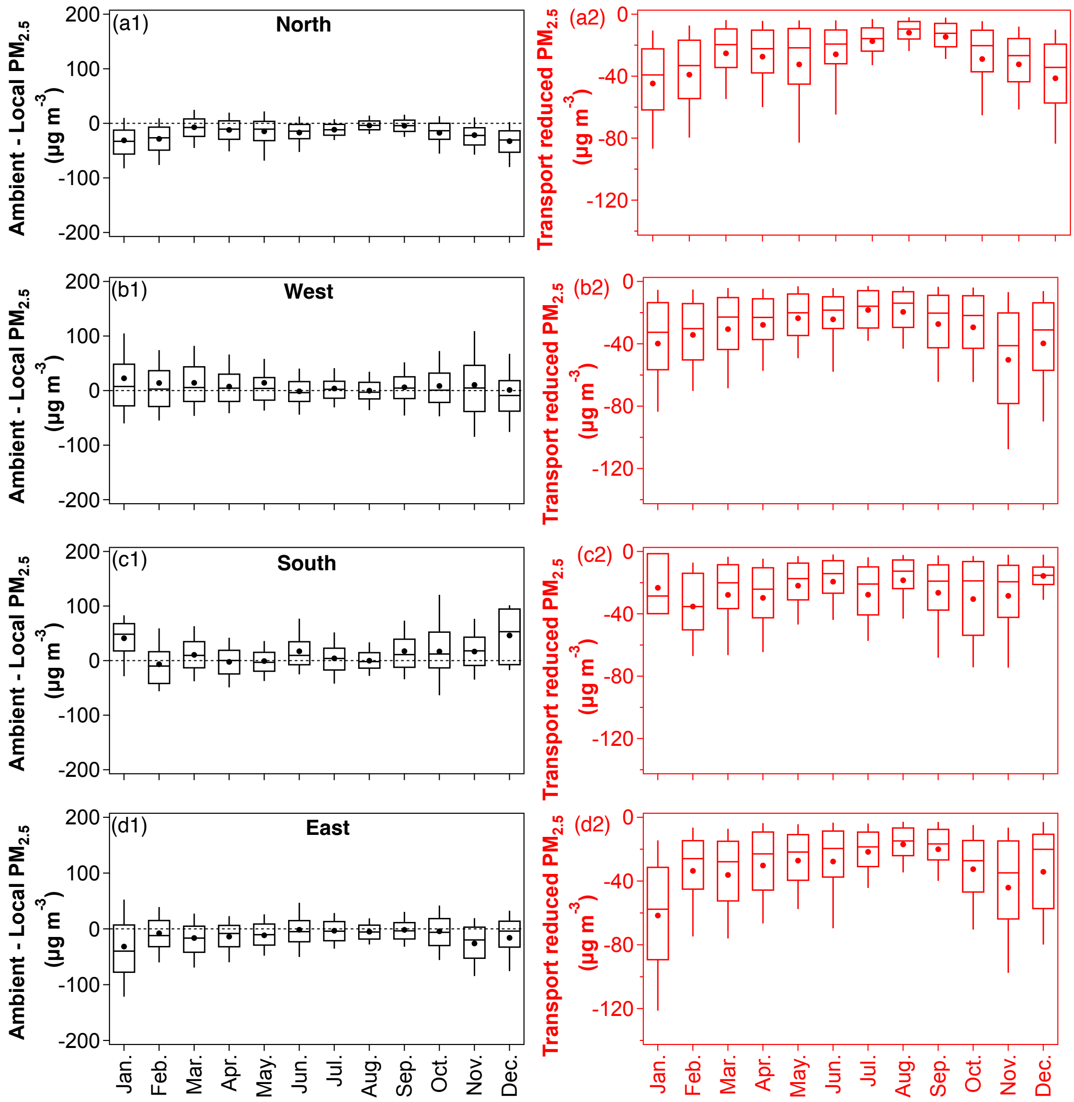

Fig. 3. Monthly variations of the difference between ambient and local $PM_{2.5}$ from the (a1) North, (b1) West, (c1) South, and (d1) East regions. Right panels show monthly variations of $PM_{2.5}$ reductions caused by regional transport for the corresponding source regions in the left panels. The upper and lower boundaries represent the 75th and 25th percentiles, respectively, while the solid origin represents the average value.

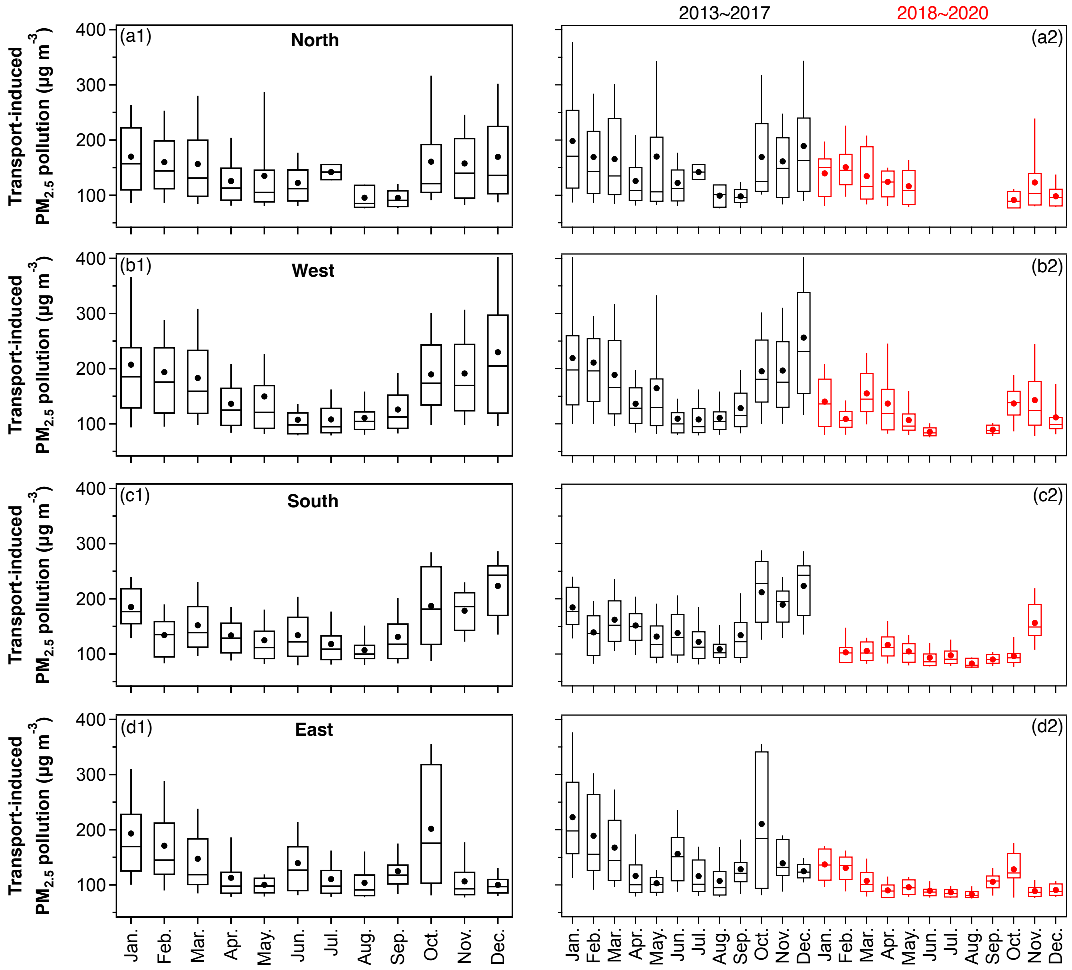

Fig. 4. Monthly variations of transport-induced $PM_{2.5}$ pollution (ambient $PM_{2.5}$ exceeding local $PM_{2.5}$ and 75 $\mu g\ m^{-3}$) from the (a1) North, (b1) West, (c1) South, and (d1) East regions. Right panels show monthly variations of transport-induced $PM_{2.5}$ pollution before (black) and after (red) 2017 for the corresponding source regions in the left panels. The upper and lower boundaries represent the 75th and 25th percentiles, respectively, while the solid origin represents the average result.

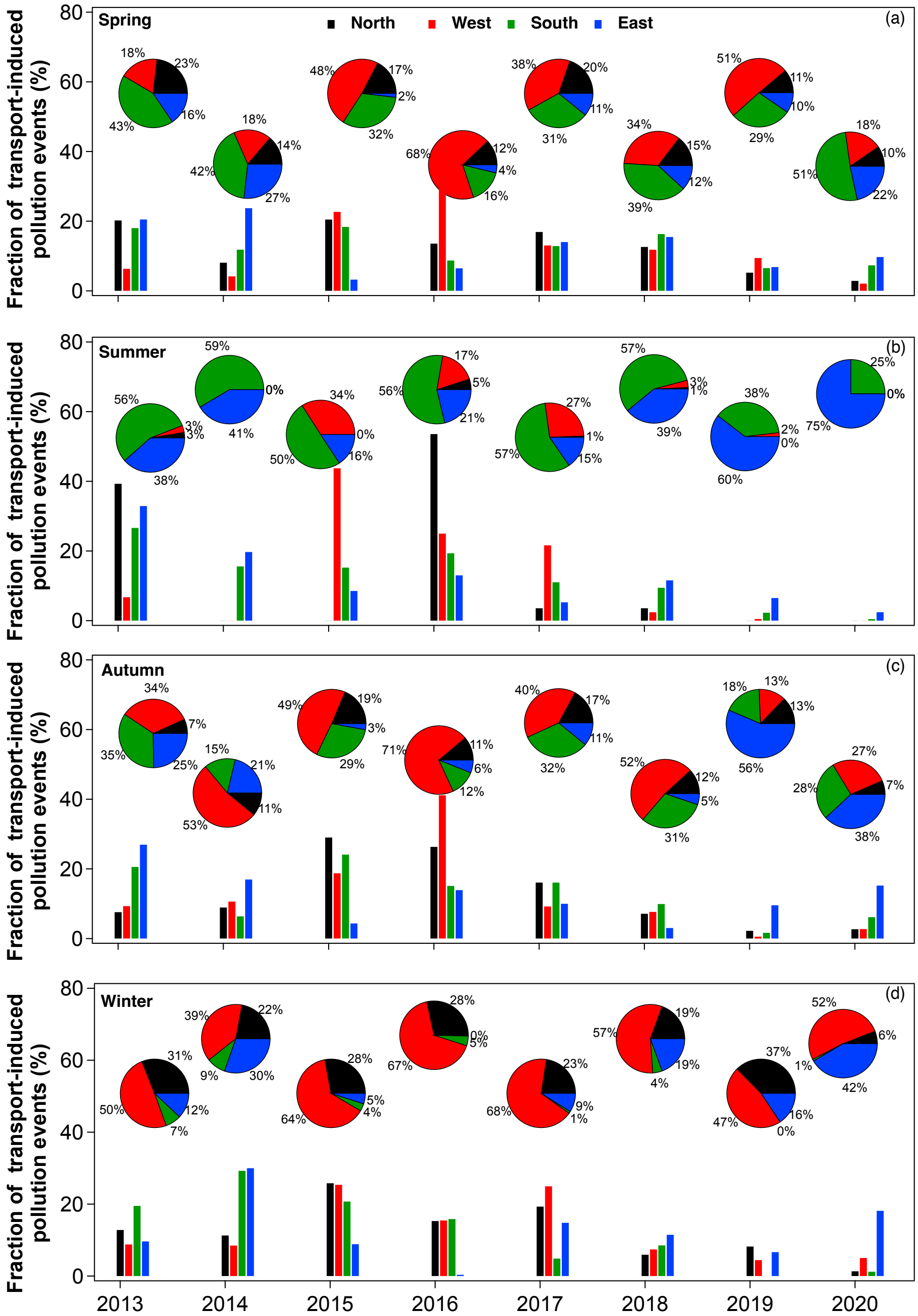

Fig. 5. Histograms depict the annual fraction of transport-induced pollution events in each direction relative to the total number of occurrences from 2013 to 2020 during (a) spring, (b) summer, (c) autumn, and (d) winter. Pie charts illustrate the proportion of transport-induced pollution events in each direction for each year within the corresponding seasons.