Title Page

https://doi.org/10.1016/j.scitotenv.2022.156740

Slower than expected reduction in annual $PM_{2.5}$ in Xi'an revealed by machine learning-based meteorological normalization

Meng Wang[1], Zhuozhi Zhang[1], Qi Yuan[1], Xinwei Li[1], Shuwen Han[1], Yuethang Lam[1], Long Cui[2], Yu Huang[2], Junji Cao[2,3], Shun-cheng Lee[1,*]

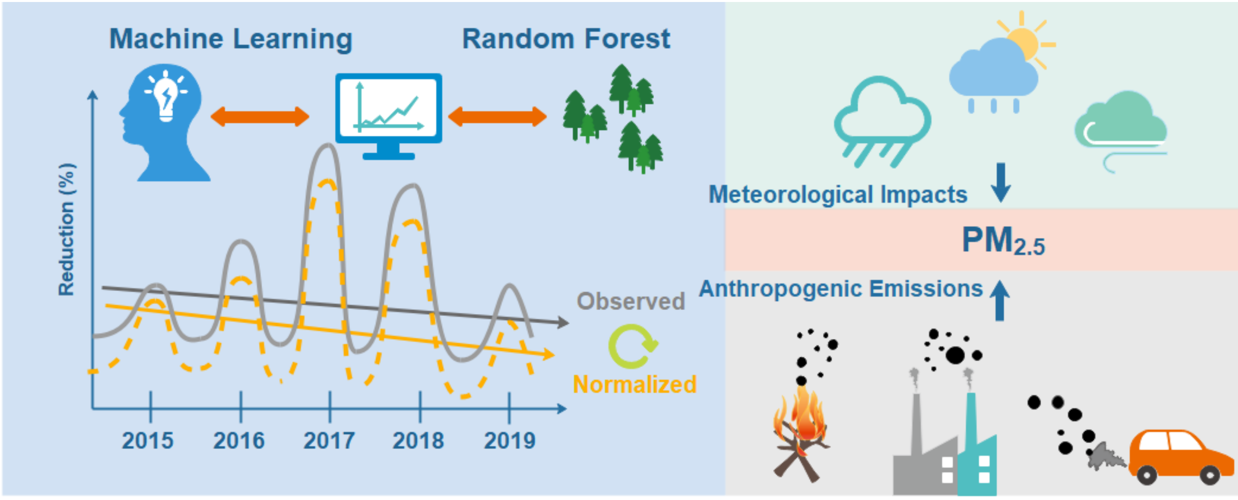
[1]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

[2]State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China

[3]Key Laboratory of Middle Atmosphere and Global Environment Observation, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Correspondence to: Shun-cheng Lee (shun-cheng.lee@polyu.edu.hk).

Graphical Abstract



Highlights

- Trend analysis of $PM_{2.5}$ over multiple years is complicated due to the impact of meteorology.
- Meteorological normalization was performed using the machine learning algorithm.
- Real trend in $PM_{2.5}$ in a polluted northwest city was revealed after meteorological normalization.
- Reduction rate in the normalized $PM_{2.5}$ over the 5 years was slower than the observed ones.
- Insights into the photochemical and aqueous phase chemistry of secondary $PM_{2.5}$ were gained.

Slower than expected reduction in annual $PM_{2.5}$ in Xi'an revealed by machine learning-based meteorological normalization

Meng Wang[1], Zhuozhi Zhang[1], Qi Yuan[1], Xinwei Li[1], Shuwen Han[1], Yuethang Lam[1], Long Cui[2], Yu Huang[2], Junji Cao[2,3], Shun-cheng Lee[1, *]

[1]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

[2]State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China

[3]Key Laboratory of Middle Atmosphere and Global Environment Observation, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Correspondence to: Shun-cheng Lee (shun-cheng.lee@polyu.edu.hk).

Abstract. To evaluate the effectiveness of air pollution control policies, trend analysis of the air pollutants is often performed. However, trend analysis of air pollutants over multiple years is complicated by the fact that changes in meteorology over time can also affect the levels of air pollutants in addition to changes in emissions or atmospheric chemistry. To decouple the meteorological effect, this study performed a trend analysis of the hourly fine particulate matter ($PM_{2.5}$) observed at an urban background site in Xi'an city over 5 years from 2015 to 2019 using the machine learning algorithm. As a novel way of meteorological normalization, the meteorological parameters were used as constant input for 5 consecutive years. In this way, the impact of meteorological parameters was excluded, providing insights into the "real" changes in $PM_{2.5}$ due to changes in emission strength or atmospheric chemistry. After meteorological normalization, a decreasing trend of −3.3% year^{-1} (−1.9 μg m^{-3} year^{-1}) in $PM_{2.5}$ was seen, instead of −4.4% year^{-1} from direct $PM_{2.5}$ observation. Assuming the rate of −1.9 μg m^{-3} year^{-1} were kept constant for the next few decades in Xi'an, it would take approximately 25 years (in the year 2045) to reduce the annual $PM_{2.5}$ level to 5 μg m^{-3}, the new guideline value from World Health Organization. We also show that $PM_{2.5}$ is primarily associated with anthropogenic emissions, which, underwent aqueous phase chemistry in winter and photochemical oxidation in summer as suggested by partial dependence of RH and O_x in different seasons. Therefore, reducing the anthropogenic secondary aerosol precursors at a higher rate, such as NO_x and VOCs is expected to reduce the particulate pollution in this region more effectively than the current −3.3% year^{-1} found in this study.

***Keywords*:** Particulate matter; Secondary aerosol; Theil-Sen estimator; Random forest; Aqueous phase chemistry

1 Introduction

Atmospheric particulate matter with a diameter of less than 2.5 μm ($PM_{2.5}$) is associated with adverse health effects and plays a key role in climate change (Cai et al., 2017; Daellenbach et al., 2020; Wu et al., 2022). Globally, atmospheric $PM_{2.5}$ is causing millions of premature deaths every year (Burnett et al., 2018; Cohen et al., 2017; Lelieveld et al., 2015). In particular, exposure to high levels of $PM_{2.5}$ is associated with a high risk of cardiovascular and respiratory disease (Brehmer et al., 2019; Lyu et al., 2018; Yu et al., 2019). The World Health Organization (WHO) recommends an annual $PM_{2.5}$ level of 10 μg m^{-3} (or more recently 5 μg m^{-3}) not to be exceeded (WHO, 2006; WHO, 2021). However, it is noted that there is no safe level of $PM_{2.5}$ below which no adverse health effects would be anticipated (WHO, 2006).

$PM_{2.5}$ can be directly emitted from sources of e.g., traffic, industry, and coal combustion; it can also be formed from the oxidation of its precursor gases of e.g., NO_x, SO_2, volatile organic compounds (VOCs) (Fuzzi et al., 2015; Shrivastava et al., 2017; Zhang et al., 2015), termed as primary and secondary $PM_{2.5}$, accordingly. To evaluate the effectiveness of air pollution control policies and to further inform policy development, trend analysis of the observed $PM_{2.5}$ in the ambient environment upon changes in emission and atmospheric chemistry over time is important, especially on a long-term basis e.g., years to decades. However, trend analysis of $PM_{2.5}$ over multiple years is complicated because changes in meteorology can drive the changes in the observed $PM_{2.5}$ in addition to changes in emissions or atmospheric chemistry. Therefore, it is essential to decouple the meteorological impact from the observed $PM_{2.5}$ to see the real changes caused by emission over time with statistical significance.

To confirm the changes in pollutant concentration over multiple years with statistical significance, a process called meteorological normalization was proposed by Grange et al. (2018) using the random forest-based machine learning algorithm. The random forest model is computationally efficient and can well predict the $PM_{2.5}$ based on the meteorological parameters (Vu et al., 2019; Zhan et al., 2022; Zhou et al., 2022). By eliminating the effect of meteorological parameters (i.e., after meteorological normalization), insights into the real changes due to emission strength over time can be gained (Grange and Carslaw, 2019; Grange et al., 2021; Grange et al., 2017). Moreover, Qin et al. (2022) show the nonlinear effect of atmospheric variables on the primary and secondary organic aerosol can be well captured by the random forest model. Using the partial dependence algorithm, the emission sources and formation process of $PM_{2.5}$ can be revealed in a complex urban environment (Qin et al., 2022).

In China, $PM_{2.5}$ pollution is particularly serious due to rapid economic development, industrialization, and urbanization (Lu et al., 2013; Zhang et al., 2012). To tackle air pollution, many measures have been implemented e.g., the 5-year Clean Air Action Plan and the blue-sky action (Cheng et al., 2019; Wang et al., 2014; Yang et al., 2015). Despite the efforts to reduce the emission, recent studies show the annual $PM_{2.5}$ concentration in Northern China is still far exceeding the WHO guideline values (Chen et al., 2019; Cheng et al., 2019; Vu et al., 2019), highlighting the challenges to improve the air quality in China. As the largest city in northwest China and home to 13 million people as of 2021, Xi'an has suffered severe air pollution over the past decades with $PM_{2.5}$ levels typically higher than in Beijing (Dai et al., 2018; Elser et al., 2016; Huang et al., 2014; Niu et al., 2016). However, compared to Beijing, trend analysis of $PM_{2.5}$ in this highly polluted city is lacking, limiting our understanding of the most recent changes in the evolution of $PM_{2.5}$ over time. In particular, while it is widely acknowledged that 5-year Clean Air Action Plan is contributing to the reduction of $PM_{2.5}$ levels in Beijing (Cheng et al., 2019; Vu et al., 2019), it is unknown if the clean air action is working in Xi'an as significantly.

In this study, trend analysis of the hourly $PM_{2.5}$ over 5 years from 2015 to 2019 in Xi'an was performed using the random forest model. The random forest model was used to predict the $PM_{2.5}$ using the meteorological parameters as the model input. Through the comparison of trend analysis before and after meteorological normalization, the effect of meteorological on trend estimates is revealed. Using the partial dependence algorithm, the nonlinear effects of atmospheric variables and gaseous pollutant on $PM_{2.5}$ was evaluated. Finally, implications from trend analysis of $PM_{2.5}$ over the 5 years in Xi'an are discussed.

2 Method

2.1 Data source

Five years of air quality data (from 2015 to 2019) of the hourly $PM_{2.5}$, NO_2, SO_2, O_3 and CO at three national air quality monitoring stations in Xi'an were downloaded from the China National Environmental Monitoring Network website (https://www.cnemc.cn/;last access: February 1, 2022). The three sampling sites are all within the urban Xi’an, specifically in three different districts in Xi'an, with GXXQ in Gaoxin District, XZ in Yanta District, and LTQ in Lintong District (Fig. S1). The distance between GXXQ and LTQ sampling site is approximately 40 km, while it is 5 km between GXXQ and XZ (Fig. S1). With such large spatial coverage, the air quality data recorded at the three sampling sites can represent the overall air quality in Xi'an city, one of the most polluted cities in China.

Hourly meteorological data including wind speed, wind direction, temperature, relative humidity (RH) recorded at Xi'an Xianyang International Airport were downloaded using the “worldMet” R package (Carslaw, 2017). Planetary boundary layer (PBL) height and atmospheric pressure were obtained from the reanalysis data at 100 m above ground level at the sampling site of GXXQ using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model (Draxler and Rolph, 2003), developed by the National Oceanic and Atmospheric Administration (NOAA). Data were analyzed in RStudio with a series of packages, including “openair”, “normalweather”, and “ggplot2” (Carslaw and Ropkins, 2012; Grange et al., 2018; Vu et al., 2019).

2.2 Random Forest modelling

2.2.1 Building the Random Forest model

A decision-tree-based random forest model was developed to understand the trend of the observed $PM_{2.5}$ over the 5 years and to gain insights into the formation pathways of $PM_{2.5}$. Specifically, the random forest model was built to derive the relationship between $PM_{2.5}$ and its predictor features including time variables (date_unix (number of seconds since 1 January 1970), day of the year (day_julian), weekday, and hour of the day), meteorological parameters (wind speed, wind direction, temperature, relative humidity (RH), PBL, and pressure). The time variables act as proxies for emission strength as they vary in time and season.

In the RF model, the whole dataset was randomly divided into a training dataset to build the model and a testing dataset to test the model performance. The training dataset was comprised of 80% of the whole dataset, with the testing data (20%) used to validate the models once the forest had been grown. The number of the independent/explanatory variables used to grow a tree was set to three, while the minimum nod-size was set to five, following (Grange et al., 2018). The number of trees within a forest was set to 300. The RF model was built using the latest “rmweather” R package developed by Grange et al. (2018).

2.2.2 Meteorological normalization

$PM_{2.5}$ can be meteorologically normalized by repeatedly (1000 times) re-sampling and predicting using the random models as detailed by Grange et al. (2018). Briefly, $PM_{2.5}$ at a specific measured time point with randomly resampled explanatory variables (except for date_unix) is predicted 1000 times and averaged. For every prediction, the explanatory variables including the time variables (excluding the date_unix variable) and meteorological parameters were randomly selected from the original observation dataset and were subsequently fed to the RF model to predict $PM_{2.5}$ at that particular time point. This is repeated 1000 times, and the 1000 predictions were then averaged, representing "average" meteorological conditions and hence, was regarded as the meteorologically normalized trend. In other words, the meteorological normalized $PM_{2.5}$ (in μg m^{-3}) can be thought of as concentrations in "average" or invariant weather conditions. Because the time variables of the hour, weekday, day of the year are also included for normalization, it is not straightforward to investigate the hourly, weekday, seasonal for a comparison with the trend of the observed values.

In this study, the meteorological parameters in 2015 were used as the input to predict the $PM_{2.5}$ concentrations in 2016, 2017, 2018, and 2019. In other words, the predicted $PM_{2.5}$ in 2016, 2017, 2018, 2019 were the expected $PM_{2.5}$ concentrations under 2015 meteorological conditions. In this way, the predicted $PM_{2.5}$ can be directly compared with the observed $PM_{2.5}$ in terms of hourly, weekday, seasonal variations. Note that only the meteorological parameters (i.e., wind speed, wind direction, temperature, pressure, and PBL) were re-sampled, while the time variables were unchanged. With the predicted $PM_{2.5}$ under the same meteorological conditions from 2015-2019, the behavior of the $PM_{2.5}$ trend due to the changes in emissions or atmospheric chemistry can be revealed.

2.2.3 Partial dependence algorithm

The partial dependence algorithm was applied to assess the nonlinear effect of atmospheric variables, including physical and chemical processes, on the measured $PM_{2.5}$ (Grange and Carslaw, 2019; Grange et al., 2018). The partial dependence algorithm calculated the dependence between the $PM_{2.5}$ and the target atmospheric variables while holding other variables constant at their averages. By targeting all variables one by one, the partial dependence of $PM_{2.5}$ on all considered atmospheric variables was calculated.

In this study, the atmospheric variables used as model input included meteorological parameters and gas pollutants. Specifically, the meteorological parameters were RH and temperature which are key atmospheric variables that can influence the physical and chemical processes of $PM_{2.5}$. For example, high RH may promote aqueous phase chemistry (Duan et al., 2020), while the high temperature may induce high biogenic VOC emissions in summer, key precursor gases for secondary aerosol. Gas pollutants include CO, SO_2, NO_2, O_3, as well as O_x (NO_2 + O_3). CO and SO_2 are indicators of primary emissions, while, O_x is a good surrogate of the oxidizing capability of the atmosphere (Lin et al., 2020). Note that although CO and SO_2 are primary emissions, they are not necessarily local since they can be transported from upwind regions to the receptor sites. The partial dependence algorithm is provided in the "rmweather" package (Grange et al., 2018) in R (version 4.1.2)

2.3 Trend analysis using Theil-Sen estimator

The Theil-Sen regression methodology was applied to investigate the long-term trend of $PM_{2.5}$ before and after the meteorological normalization. The Theil-Sen approach is commonly used for long-term trend analysis and has been detailed in Grange et al. (2018) and Vu et al. (2019). Briefly, the Theil-Sen regression approach accounted for autocorrelation and was used at the 95% confidence level to indicate a significant trend (Grange et al., 2018). The Theil-Sen approach computed the slopes of all possible pairs of $PM_{2.5}$ and took the median values of the slopes, resulting in more conservative confidence intervals for $PM_{2.5}$ trend analysis. The Theil-Sen functions are provided in the “openair” package in R (version 4.1.2) (Carslaw and Ropkins, 2012).

3 Results and Discussion

3.1 Ambient $PM_{2.5}$ in Xi'an from 2015 to 2019

Figure 1 shows the daily averaged time series of $PM_{2.5}$ over the five years from 2015 to 2019 at the three different sites (i.e., LTQ, XZ, and GXXQ) in Xi'an. The time series of $PM_{2.5}$ at the three sites were very similar with elevated concentrations in winter (spiking over 400 μg m^{-3}) and relatively reduced concentrations in summer (< 100 μg m^{-3}). Averaged over the five years, $PM_{2.5}$ was 65.1 ± 59.9 (SD) μg m^{-3} at GXXQ, while it was 62.2 ± 61.2μg m^{-3} and 59.3 ± 58.6 μg m^{-3} at XZ and LTQ, respectively (Table S1). Despite the large distance between the sampling sites (up to 40 km; Figure S1), the time series of $PM_{2.5}$ at the three sites were highly correlated with correlation coefficient $r > 0.85$ (p-value < 0.01) and slopes close to unity. The good correlation for the observed $PM_{2.5}$ at the three sampling sites suggests the observed $PM_{2.5}$ were due to common pollution sources, simultaneously impacting the air quality over a large area in Xi'an with a diameter of at least 40 km. Due to the similar trend in time series and the slightly high concentration observed at GXXQ, below we focus on the discussion on the air quality data at GXXQ.

In terms of annual mean concentration, the $PM_{2.5}$ at GXXQ was 63.6 μg m^{-3} in 2015. It increased to 74.8 μg m^{-3} in 2017 then dropped to 58.8 μg m^{-3} in 2019 (Table S1). Compared to China’s national ambient air quality standard (NAAQS-II) of 35 μg m^{-3} and the new WHO guideline of 5 μg m^{-3} (WHO, 2021), the annual mean $PM_{2.5}$ concentration in Xi'an was approximately substantially (2-7 times) higher, highlighting the poor air quality in this city. Moreover, compared to the trend of $PM_{2.5}$ in Beijing (Vu et al., 2019), which showed a decreasing trend from 88 μg m^{-3} in 2013 to 58 μg m^{-3} in 2017, the $PM_{2.5}$ trend observed in Xi'an is more complicated since the annual $PM_{2.5}$ concentration increased in 2017 then started decreased afterward. In particular, the number of haze days (defined as daily $PM_{2.5}$ > 75 μg m^{-3}) was 90 days (i.e., ~25% of the year or 1 in 4 days; Table S2) in 2015. It increased to 112 days in 2017 then dropped to 86 days in 2019 (Table S2). Most of the haze days occurred in winter, with the average $PM_{2.5}$ concentrations in the range of 67.3-143 μg m^{-3} in winter (Table S3), roughly three times higher than in summer (24.5-38.6 μg m^{-3}).

3.2 Predicted $PM_{2.5}$ in a good agreement with the observed $PM_{2.5}$ over 5 years

A decision-tree-based random forest model was trained for the observed $PM_{2.5}$ with the independent variables including time variables and meteorological parameters as the model input (see the Method section). During the model building, 80% of the dataset was randomly selected as the training dataset, with the rest 20% as the testing dataset. For the training dataset, the predicted $PM_{2.5}$ was well correlated with the observed $PM_{2.5}$ with R^2 of 0.99 and slope of 0.93 (Figure 2), while for the testing dataset, the model reproduced the observed $PM_{2.5}$ reasonably well with R^2 of 0.93 and slope of 0.84. The slope of 0.84-0.93 for the testing dataset suggested the model tended to underestimate the $PM_{2.5}$ by 7-16%. Nevertheless, the high R^2 values (0.93-0.99) for both the training and testing dataset suggest the random forest grown in this study had a strong explanatory ability for $PM_{2.5}$.

The good performance of the random forest model was partly due to the strong seasonality of the $PM_{2.5}$

which was well captured by the model. Specifically, the time variable (i.e., day of the year or day_Julian (1-365)) was the most important variable for $PM_{2.5}$ explanation in the random forest model (Figure S2). Partial dependence on the time variable of day_julian shows the elevated $PM_{2.5}$ concentrations (> 75 μg m^{-3}) were associated with day 1-50 and day 300-365 in the year (Figure S3), consistent with the fact that haze pollution occurred mostly in winter. In contrast, the time variable of weekday and hour were of less importance. Partial dependence plots on the weekday do not present a clear weekday/weekends difference (Figure S3). Given that traffic is usually heavier during weekdays than weekends, the lack of weekday and weekends pattern suggests traffic was not the major source of $PM_{2.5}$. Consistently, the partial dependence on the time variable of hour shows no rush hour peaks (Figure S3). Instead, elevated $PM_{2.5}$ concentrations were found to be occurring mostly in the night till the next morning (Figure S3).

Among the meteorological parameters, temperature was the most important parameter, followed by pressure, relative humidity, boundary layer height, wind speed and wind direction (Figure S2). In particular, the low temperature (< 10 ºC), low pressure (< 880 hpa), high relative humidity (50-90%), low boundary layer height (< 500 m), low wind speed (< 3 m s^{-1}) in north-eastly wind were associated with high $PM_{2.5}$ concentrations (Figure S3). These meteorological parameters created a stable atmosphere with poor dispersion conditions, causing the build-up of $PM_{2.5}$. Note that at higher relative humidity (95-100%), precipitation events likely caused the wet deposition of $PM_{2.5}$, leading to lower concentrations as a result (Figure S3). Therefore, the random forest model captured the general predictions and processes that were associated with the ambient $PM_{2.5}$ concentrations, confirming the strong explanatory power of the model.

3.3 Trend analysis before and after meteorological normalization

Figure 3 shows the yearly averaged $PM_{2.5}$ concentration before and after meteorological normalization. It shows that $PM_{2.5}$ concentration in 2019 would have been higher if under the 2015 meteorological conditions, while $PM_{2.5}$ concentrations in 2017 would have been lower if under the same 2015 meteorological conditions. Specifically, the observed $PM_{2.5}$ concentrations (i.e., before meteorological normalization) were 63.6 μg m^{-3}, 66.5 μg m^{-3}, 74.8 μg m^{-3}, 61.2 μg m^{-3}, 58.8 μg m^{-3}, in 2015, 2016, 2017, 2018, 2019, respectively. After meteorological normalization, the predicted $PM_{2.5}$ concentrations from 2015 to 2019 were 63.4 μg m^{-3}, 66.8 μg m^{-3}, 72.3 μg m^{-3}, 64.1 μg m^{-3}, 60.9 μg m^{-3}, respectively. This can be translated to percentages differences of −0.3%, 0.4%, −3.3%, 4.7%, 3.6% by comparing the predicted and observed $PM_{2.5}$. The percentage differences may appear small (from −3.3% to 4.7%) when compared with the observed $PM_{2.5}$. However, in terms of the yearly $PM_{2.5}$ trend analysis which is usually on the scale of 1-10% (Vu et al., 2019), the changes in $PM_{2.5}$ due to different meteorological conditions may have a big impact. Below, we discuss the effect of meteorological normalization on $PM_{2.5}$ trend analysis.

Figure 4 shows the trend analysis of monthly averaged $PM_{2.5}$ before and after meteorological normalization using the same Theil-Sen algorithm (see Sect. 2.4). The temporal variations of the monthly average $PM_{2.5}$ for both cases do not show a smooth trend from 2016 to 2019 because of the spikes during pollution events in cold seasons, consistent with the daily average $PM_{2.5}$ as shown in Figure 1. Using the Theil-Sen estimator, the observed $PM_{2.5}$ (before meteorological normalization) shows a trend of -4.4% $year^{-1}$ or −2.6 μg m^{-3}, while it shows a less negative trend of −3.3% $year^{-1}$ (−1.9 μg m^{-3}) after meteorological normalization. However, both show a large range in terms of 95% confidence level, from −10.76% $year^{-1}$ to 6.1% $year^{-1}$ for the observed $PM_{2.5}$ and slightly positive values from −9.41% $year^{-1}$ to 6.7% $year^{-1}$ for the meteorological normalized $PM_{2.5}$. The large range of the 95% confidence level was due to the fact that the pollution events in cold seasons do not appear abating in terms of the magnitude of the $PM_{2.5}$ levels and the duration of the pollution. Nevertheless, compared to the observed $PM_{2.5}$ trend, the slightly positive trend of $PM_{2.5}$ after meteorological normalization results suggest that the effect of emission reduction was contributing less to the improvement of air quality in Xi'an. This is in great contrast to the findings in Beijing, where emission reductions were found to cause a larger reduction in $PM_{2.5}$ after meteorological normalization although with different normalization methodology over different years (2013-2017; (Vu et al., 2019)).

3.4 Formation process of haze pollution

As discussed above, the most severe pollution events occurred in the winter months (December, January, and February) over the 5 years from 2016 to 2019. To gain insights into the formation processes of haze pollution in winter, gas pollutants of CO, NO_2, O_3, SO_2, and O_x were fed into the random forest in addition to the meteorological parameters of RH, temperature, wind speed, and wind direction. As a comparison, a similar analysis was also performed during the summer months (June, July, and August). Note that because we focused on only one season, time variables were not considered. Additionally, a multi-linear regression model was also performed between $PM_{2.5}$ and these gas pollutants. The results show the correlation determination R^2 for the random forest model (0.64-0.71) was significantly higher than for the multi-linear regression model (< 0.4; Table S4). Therefore, the random forest model can provide higher accuracies of the $PM_{2.5}$ prediction than the multi-linear regression model.

Figure 5 shows the importance parameters during the model training process for the winter and summer $PM_{2.5}$. In both winter and summer, CO is the most important variable in explaining the observed $PM_{2.5}$ (Figure 5). Given that CO is the by-product of incomplete combustion, the strong importance of

CO in explaining $PM_{2.5}$ suggests $PM_{2.5}$ was primarily associated with anthropogenic combustion sources including both direct emission and/or secondary formation from anthropogenic precursor gases. However, as discussed above, the good time series correlation between the three sampling sites (despite a distance of 40 km) suggests the observed $PM_{2.5}$ were regionally relevant rather than local pollution events, and, therefore, the observed $PM_{2.5}$ was likely associated with the secondary formation during transport. Indeed, recent studies of $PM_{2.5}$ source apportionment highlight secondary aerosols are the major component of $PM_{2.5}$ instead of primary emission in Xi'an (Duan et al., 2020; Elser et al., 2016; Zhong et al., 2020).

Figure 5 also shows, while RH was the second most important parameter in winter, O_x (NO_2 + O_3) was the second most important variable in summer. The partial dependence plot shows that high RH was associated with high $PM_{2.5}$ in winter (Figure 6), while high O_x was associated with high $PM_{2.5}$ in summer. Recent studies show RH can promote the aqueous formation of secondary aerosol including sulfate and oxygenated organic aerosol, while these secondary aerosols together often contribute over half of the $PM_{2.5}$ mass (Elser et al., 2016; Zhong et al., 2020). In this study, the aqueous phase chemistry is reflected by the high importance of RH in winter, with the overall $PM_{2.5}$ showing a positive response to RH in winter. In contrast, O_x is a good indicator of photochemical chemistry, which is more important in summer than in winter as reflected by the importance of O_x in summer (Figure 5), showing a positive response to O_x (Figure 6). As a comparison, O_x in winter was the least important gaseous variable in winter, suggesting photochemical chemistry was less significant than RH-promoted aqueous phase chemistry. In summer, RH was still the fourth important variable after O_3, implying aqueous phase chemistry could also be the major pathway for secondary aerosol formation. Consistently, Duan et al. (2020) shows a large formation of secondary aerosol formation during fog-rain days in summer Xi'an.

4 Discussion

Using the random forest model, we show that the 5-year hourly $PM_{2.5}$ measured at the suburban site in Xi'an from 2015-2019 was reproduced well by feeding the meteorological parameters and time variables into the model. Meteorological parameters can affect the dispersion conditions and/or atmospheric chemistry of the ambient $PM_{2.5}$, while the time variables act as proxies for emission strength as they vary in terms of hour, day, season, and year. Assuming the meteorological parameters were the same throughout the 5 years (i.e., normalization), we can exclude the impact of meteorological parameters, providing insights into the "real" changes in $PM_{2.5}$ due to changes in emission strength or atmospheric chemistry. After meteorological normalization, we show that the $PM_{2.5}$ concentration in 2019 would have been higher, while $PM_{2.5}$ concentrations in 2017 would have been lower if under the same meteorological conditions as in 2015. As a result, a decreasing trend of −3.3% year^{-1} in $PM_{2.5}$ after meteorological normalization was seen, instead of −4.4% from direct $PM_{2.5}$ observation. The "real" decreasing rate of −3.3% year^{-1} for $PM_{2.5}$ in Xi'an was roughly half of the values (−7.8% year^{-1}) reported in Beijing over the year of 2013-2017 (Vu et al., 2019). Assuming the rate of −3.3% year^{-1} or 1.9 μg m^{-3} year^{-1} were kept constant for the next few decades in Xi'an, it would take approximately 25 years (in the year 2045) to reduce the yearly $PM_{2.5}$ concentration to 10 μg m^{-3}, the guideline value from WHO. Therefore, more efforts need to be taken to reduce the $PM_{2.5}$ pollution in this inland city, which is the large northwestern city in China, home to over 10 million in northwest China.

We also show that the non-linear effect of atmospheric variables on $PM_{2.5}$ can be captured by the random forest model as opposed to the multi-linear regression. Different from the multi-linear regression model, the random forest model also provides insights into the relative importance of the atmospheric variable. In particular, we show that in both winter and summer, CO is the most important variable, suggesting the observed $PM_{2.5}$ is primarily associated with anthropogenic emissions, which, undergoes aqueous phase chemistry in winter and photochemical oxidation in summer as suggested by importance of RH and O_x, the second most important variable, accordingly, after CO. Given that the time series of $PM_{2.5}$ are well correlated at the three sampling sites, despite a distance of 40 km apart, the secondary formation pathways, which is different in different seasons, play an important role covering a large area in Xi'an. As a result of secondary formation, the difference in $PM_{2.5}$ concentration at the three sampling sites is marginal. Therefore, reducing the anthropogenic secondary aerosol precursors at a higher rate, such as NO_x and VOCs is expected to reduce the particulate pollution in this region at a faster pace than the current −3.3% year^{-1} found in this study.

5 Conclusion

In this study, trend analysis of the hourly fine particulate matter ($PM_{2.5}$) observed at an urban background site in Xi'an city over 5 years from 2015 to 2019 was performed using the machine learning algorithm - random forest model. To decouple the meteorological effect, the meteorological parameters were assumed the same throughout the 5 years. In this way, the impact of meteorological parameters was excluded, providing insights into the “real” changes in $PM_{2.5}$ due to changes in emission strength or atmospheric chemistry over 5 years. After meteorological normalization, the “real” decreasing trend of −3.3% year^{-1} in $PM_{2.5}$ after meteorological normalization was roughly 30% higher than the trend of −4.4% year^{-1} from direct $PM_{2.5}$ observation. Therefore, meteorological normalization made the decreasing trend of $PM_{2.5}$ less significant. The “real” decreasing rate of −3.3% year^{-1} for $PM_{2.5}$ in Xi'an was roughly half of the values (−7.8% year^{-1}) reported in Beijing over the year of 2013-2017, suggesting the air quality control measures were less effective in this region. To take the decreasing trend into context, we assumed the rate of −3.3% $year^{-1}$ or 1.9 μg m^{-3} $year^{-1}$ were kept constant for the next few decades in Xi'an. Then, it would take 25 years (in the year 2045) to reduce the yearly $PM_{2.5}$ concentration to 10 μg m^{-3}. Through relative importance analysis and partial dependence algorithm, the observed $PM_{2.5}$ was found to be primarily associated with anthropogenic emissions, which, underwent aqueous phase chemistry in winter and photochemical oxidation in summer. Therefore, reducing the anthropogenic secondary aerosol precursors at a higher rate, such as NO_x and VOCs is expected to reduce the particulate pollution in this region more efficiently. This study provides a robust trend analysis in $PM_{2.5}$ over 5 years in a highly polluted but less studied city in northwest China, providing high certainty that the real trend is less significant under the current control measures than observed, requiring stricter policies controlling the emission of precursor gases from anthropogenic activities.

Associate content

Supporting Information

Supplementary figures (Fig. S1-S2) and Table S1-S4.

Credit authorship contribution statement

Meng Wang: designed the study, conducted data analysis, prepared the manuscript with contributions from all co-authors.

Zhuozhi Zhang: Formal analysis, Writing, Review and Editing.

Qi Yuan: Formal analysis, Methodology.

Xinwei Li: Investigation, Methodology.

Shuwen Han: Validation, Investigation.

Yuethang Lam: Formal analysis.

Long Cui: Formal analysis, Investigation.

Yu Huang: Writing, Review and Editing.

Shun-cheng Lee: Writing - review and editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no conflicting interests.

Acknowledgements

This work was supported by the Environment and Conservation Fund - Environmental Research, Technology Demonstration and Conference Projects (ECF 63/2019), the RGC Theme-based Research Scheme (T24-504/17-N), the RGC Theme-based Research Scheme (T31-603/21-N).

References

Brehmer C, Lai A, Clark S, Shan M, Ni K, Ezzati M, et al. The Oxidative Potential of Personal and Household $PM_{2.5}$ in a Rural Setting in Southwestern China. Environmental Science & Technology 2019; 53: 2788-2798.

Burnett R, Chen H, Szyszkowicz M, Fann N, Hubbell B, Pope CA, et al. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences 2018; 115: 9592-9597.

Cai W, Li K, Liao H, Wang H, Wu L. Weather conditions conducive to Beijing severe haze more frequent under climate change. Nature Climate Change 2017; 7: 257.

Carslaw DC. Worldmet: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD), available at: http://github.com/davidcarslaw/ (last access: Feb 01, 2022), 2017.

Carslaw DC, Ropkins K. openair — An R package for air quality data analysis. Environmental Modelling & Software 2012; 27-28: 52-61.

Chen D, Liu Z, Ban J, Zhao P, Chen M. Retrospective analysis of 2015–2017 wintertime $PM_{2.5}$ in China: response to emission regulations and the role of meteorology. Atmospheric Chemistry and Physics 2019; 19: 7409-7427.

Cheng J, Su J, Cui T, Li X, Dong X, Sun F, et al. Dominant role of emission reduction in $PM_{2.5}$ air quality improvement in Beijing during 2013–2017: a model-based decomposition analysis. Atmospheric Chemistry and Physics 2019; 19: 6125-6146.

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 2017; 389: 1907-1918.

Daellenbach KR, Uzu G, Jiang J, Cassagnes L-E, Leni Z, Vlachou A, et al. Sources of particulate-matter air pollution and its oxidative potential in Europe. Nature 2020; 587: 414-419.

Dai Q, Bi X-H, Liu B, Li L, Ding J, Song W, et al. Chemical nature of $PM_{2.5}$ and PM_{10} in Xi'an, China: Insights into primary emissions and secondary particle formation. Environmental pollution (Barking, Essex : 1987) 2018; 240: 155-166.

Draxler RR, Rolph GD. HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) model access via NOAA ARL READY website (www. arl. noaa. gov/ready/hysplit4. html). NOAA Air

Resources Laboratory, Silver Spring. Md, 2003.

Duan J, Huang R-J, Gu Y, Lin C, Zhong H, Wang Y, et al. The formation and evolution of secondary organic aerosol during summer in Xi'an: Aqueous phase processing in fog-rain days. Science of The Total Environment 2020: 144077.

Elser M, Huang RJ, Wolf R, Slowik JG, Wang Q, Canonaco F, et al. New insights into $PM_{2.5}$ chemical composition and sources in two major cities in China during extreme haze events using aerosol mass spectrometry. Atmospheric Chemistry and Physics 2016; 16: 3207-3225.

Fuzzi S, Baltensperger U, Carslaw K, Decesari S, Denier Van Der Gon H, Facchini M, et al. Particulate matter, air quality and climate: lessons learned and future needs. Atmospheric Chemistry and Physics 2015; 15: 8217-8299.

Grange SK, Carslaw DC. Using meteorological normalisation to detect interventions in air quality time series. Science of The Total Environment 2019; 653: 578-588.

Grange SK, Carslaw DC, Lewis AC, Boleti E, Hueglin C. Random forest meteorological normalisation models for Swiss PM_{10} trend analysis. Atmospheric Chemistry and Physics 2018; 18: 6223-6239.

Grange SK, Lee JD, Drysdale WS, Lewis AC, Hueglin C, Emmenegger L, et al. COVID-19 lockdowns highlight a risk of increasing ozone pollution in European urban areas. Atmospheric Chemistry and Physics 2021; 21: 4169-4185.

Grange SK, Lewis AC, Moller SJ, Carslaw DC. Lower vehicular primary emissions of NO_2 in Europe than assumed in policy projections. Nature Geoscience 2017; 10: 914-918.

Huang R-J, Zhang Y, Bozzetti C, Ho K-F, Cao J-J, Han Y, et al. High secondary aerosol contribution to particulate pollution during haze events in China. Nature 2014; 514: 218.

Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 2015; 525: 367-371.

Lin C, Huang R-J, Xu W, Duan J, Zheng Y, Chen Q, et al. Comprehensive Source Apportionment of Submicron Aerosol in Shijiazhuang, China: Secondary Aerosol Formation and Holiday Effects. ACS Earth and Space Chemistry 2020; 4: 947-957.

Lu Q, Zheng J, Ye S, Shen X, Yuan Z, Yin S. Emission trends and source characteristics of SO_2, NOx, PM_{10} and VOCs in the Pearl River Delta region from 2000 to 2009. Atmospheric Environment 2013; 76: 11-20.

Lyu Y, Guo H, Cheng T, Li X. Particle Size Distributions of Oxidative Potential of Lung-Deposited Particles: Assessing Contributions from Quinones and Water-Soluble Metals. Environmental Science & Technology 2018; 52: 6592-6600.

Niu X, Cao J, Shen Z, Ho SSH, Tie X, Zhao S, et al. $PM_{2.5}$ from the Guanzhong Plain: Chemical composition and implications for emission reductions. Atmospheric Environment 2016; 147: 458-469.

Qin Y, Ye J, Ohno P, Liu P, Wang J, Fu P, et al. Assessing the Nonlinear Effect of Atmospheric Variables on Primary and Oxygenated Organic Aerosol Concentration Using Machine Learning. ACS Earth and Space Chemistry 2022; 6: 1059-1066.

Shrivastava M, Cappa CD, Fan J, Goldstein AH, Guenther AB, Jimenez JL, et al. Recent advances in understanding secondary organic aerosol: Implications for global climate forcing. Reviews of Geophysics 2017; 55: 509-559.

Vu TV, Shi Z, Cheng J, Zhang Q, He K, Wang S, et al. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. Atmospheric Chemistry and Physics 2019; 19: 11303-11314.

Wang S, Xing J, Zhao B, Jang C, Hao J. Effectiveness of national air pollution control policies on the air quality in metropolitan areas of China. Journal of Environmental Sciences 2014; 26: 13-22.

WHO. Air quality guidelines: global update 2005: World Health Organization, 2006.

WHO. https://www.who.int/health-topics/air-pollution#tab=tab_1 (lass acess: 01 March 2022), 2021.

Wu D, Zheng H, Li Q, Jin L, Lyu R, Ding X, et al. Toxic potency-adjusted control of air pollution for solid fuel combustion. Nature Energy 2022; 7: 194-202.

Yang Z, Wang H, Shao Z, Muncrief R. Review of Beijing's Comprehensive motor vehicle emission Control program, White Paper, available at: https://theicct.org/sites/default/files/publications/Beijing_Emission_Control_Programs_201511.pdf (last access: 01 March 2022), 2015.

Yu S, Liu W, Xu Y, Yi K, Zhou M, Tao S, et al. Characteristics and oxidative potential of atmospheric $PM_{2.5}$ in Beijing: Source apportionment and seasonal variation. Science of The Total Environment 2019; 650: 277-287.

Zhan J, Liu Y, Ma W, Zhang X, Wang X, Bi F, et al. Ozone formation sensitivity study using machine learning coupled with the reactivity of volatile organic compound species. Atmospheric Measurement and Technology 2022; 15: 1511-1520.

Zhang Q, He K, Huo H. Policy: cleaning China's air. Nature 2012; 484: 161.

Zhang R, Wang G, Guo S, Zamora ML, Ying Q, Lin Y, et al. Formation of Urban Fine Particulate Matter. Chemical Reviews 2015; 115: 3803-3855.

Zhong H, Huang R-J, Duan J, Lin C, Gu Y, Wang Y, et al. Seasonal variations in the sources of organic aerosol in Xi'an, Northwest China: The importance of biomass burning and secondary formation. Science of The Total Environment 2020: 139666.

Zhou W, Lei L, Du A, Zhang Z, Li Y, Yang Y, et al. Unexpected increases of severe haze pollution during the post COVID-19 period: effects of emissions, meteorology, and secondary production. Journal of Geophysical Research: Atmospheres 2022: e2021JD035710.

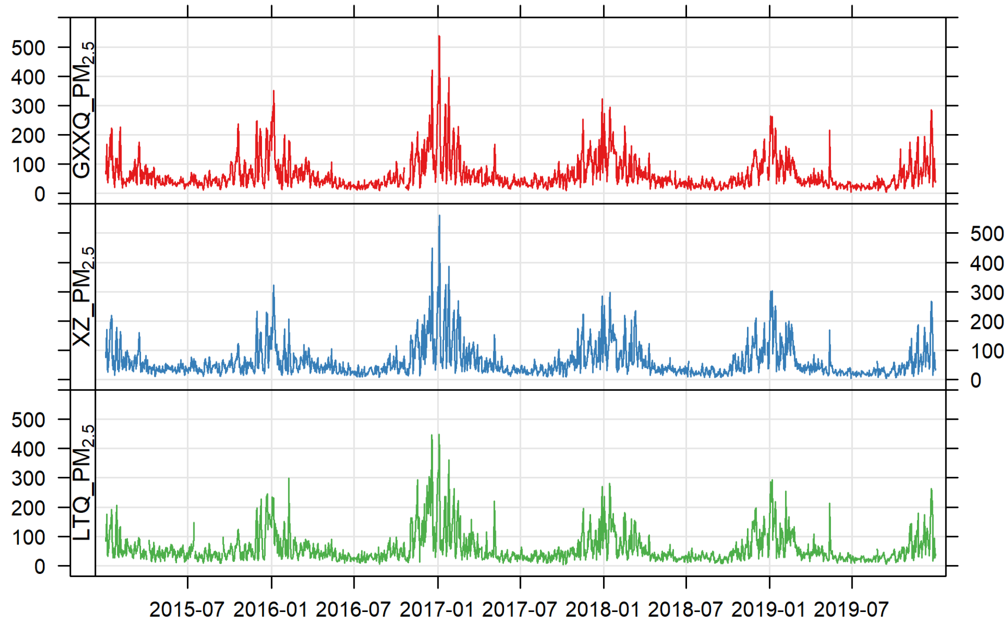


Figure 1 Time series of the daily averaged $PM_{2.5}$ (in μg m^{-3}) at the three sampling sites of GXXQ, XZ, and LTQ, with a distance of up to 40 km apart. A map of the three sampling sites is provided in Fig. S1.

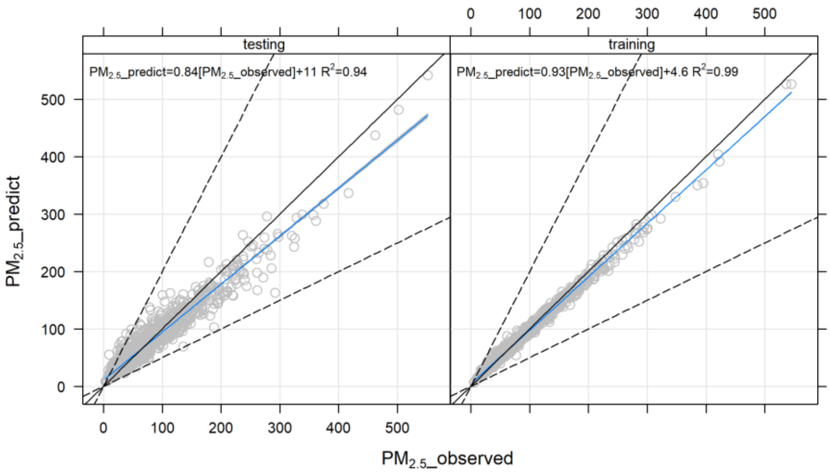


Figure 2 Scatter plots of predicted and measured $PM_{2.5}$ concentrations (in μg m^{-3}) in the train set and test set. Also shown are the linear correlation and R^2.

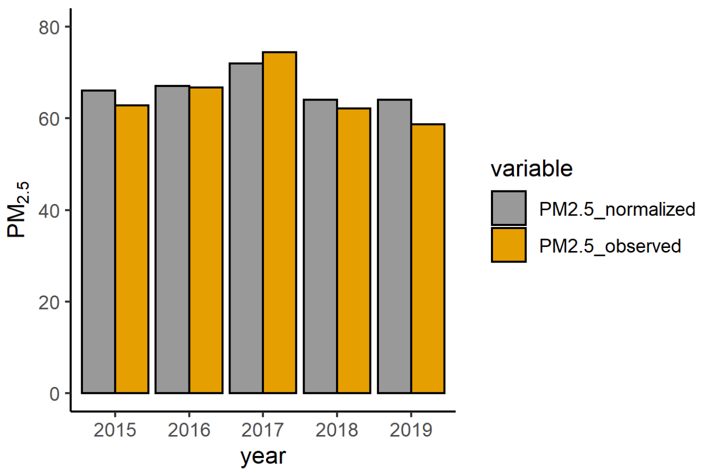


Figure 3 Annual $PM_{2.5}$ concentration (in μg m^{-3}) before (i.e., the observed $PM_{2.5}$) and after meteorological normalization (i.e., the normalized $PM_{2.5}$).

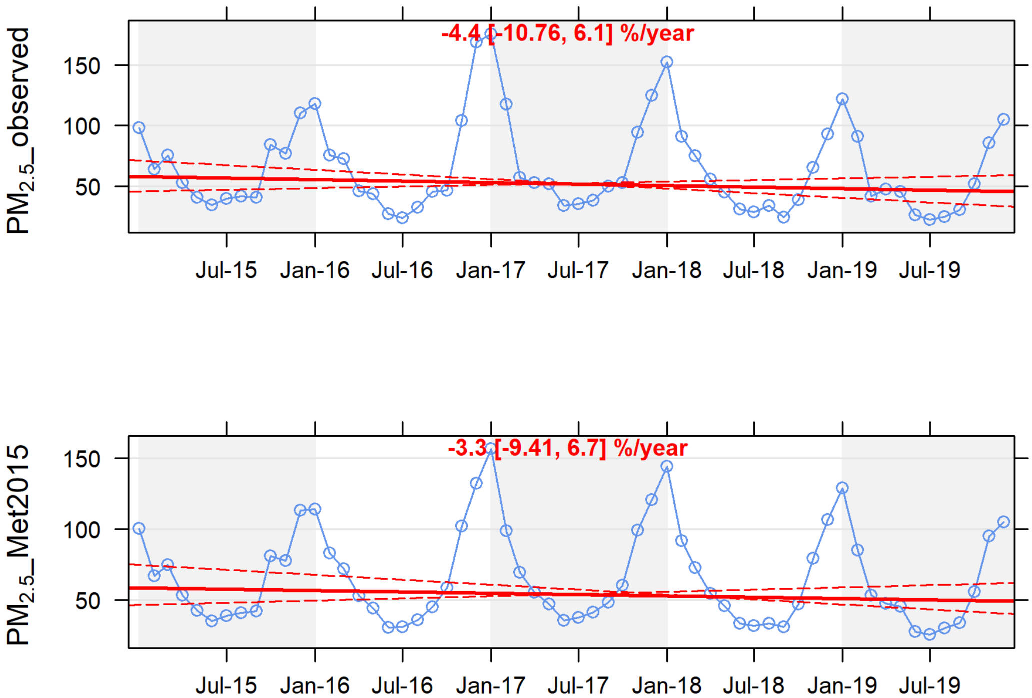


Figure 4 Monthly averaged $PM_{2.5}$ before (top panel) and after (bottom panel) meteorological normalization. The red line represents the trend analysis of $PM_{2.5}$ using the Theil-Sen estimator, with the dotted red line representing the 95% confidence level.

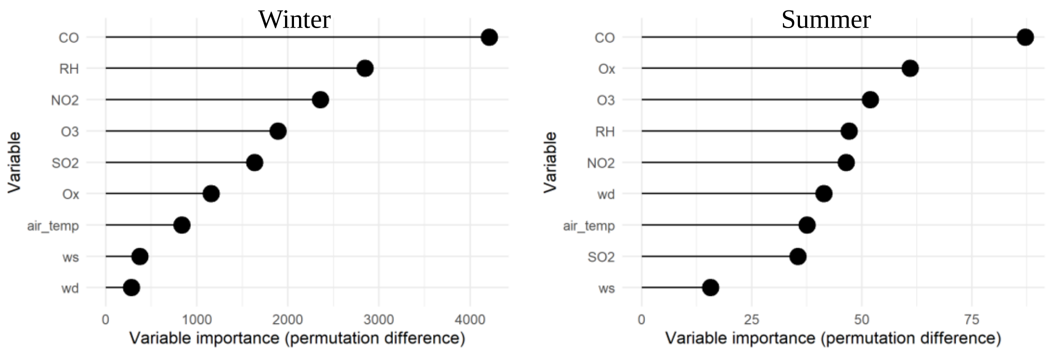


Figure 5 Variable importance for the random forest model built for the hourly $PM_{2.5}$ during winter (left panel) and summer (right panel) from 2015 to 2019. The variables include gas pollutants and meteorological parameters.

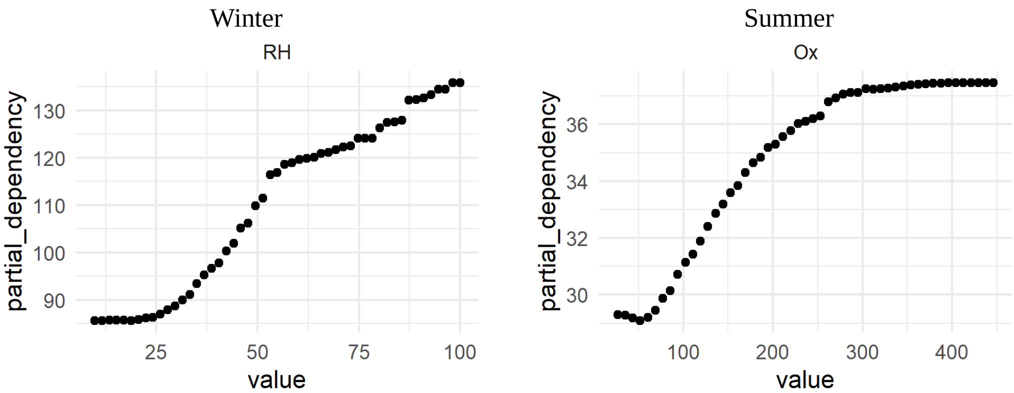


Figure 6 Partial dependence plots of the RH in winter (left panel) and O_x in summer (right panel) for the random forest model built for $PM_{2.5}$.

Credit authorship contribution statement

Meng Wang: designed the study, conducted data analysis, prepared the manuscript with contributions from all co-authors.
Zhuozhi Zhang: Formal analysis, Writing, Review and Editing.
Qi Yuan: Formal analysis, Methodology.
Xinwei Li: Investigation, Methodology.
Shuwen Han: Validation, Investigation.
Yuethang Lam: Formal analysis.
Long Cui: Formal analysis, Investigation.
Yu Huang: Writing, Review and Editing.
Shun-cheng Lee: Writing - review and editing, Funding acquisition, Supervision.