ARTICLE   OPEN

Check for updates

An intercomparison of weather normalization of PM$_{2.5}$ concentration using traditional statistical methods, machine learning, and chemistry transport models

Huang Zheng [1,2], Shaofei Kong [1,2,3 ✉], Shixian Zhai[4], Xiaoyun Sun[5], Yi Cheng[1], Liquan Yao[1], Congbo Song [6], Zhonghua Zheng [7], Zongbo Shi [8] and Roy M. Harrison [8,9 ✉]

Traditional statistical methods (TSM) and machine learning (ML) methods have been widely used to separate the effects of emissions and meteorology on air pollutant concentrations, while their performance compared to the chemistry transport model has been less fully investigated. Using the Community Multiscale Air Quality Model (CMAQ) as a reference, a series of experiments was conducted to comprehensively investigate the performance of TSM (e.g., multiple linear regression and Kolmogorov–Zurbenko filter) and ML (e.g., random forest and extreme gradient boosting) approaches in quantifying the effects of emissions and meteorology on the trends of fine particulate matter (PM$_{2.5}$) during 2013−2017. Model performance evaluation metrics suggested that the TSM and ML methods can explain the variations of PM$_{2.5}$ with the highest performance from ML. The trends of PM$_{2.5}$ showed insignificant differences ($p > 0.05$) for both the emission-related (PM$_{2.5}^{EMI}$) and meteorology-related components between TSM, ML, and CMAQ modeling results. PM$_{2.5}^{EMI}$ estimated from ML showed the least difference to that from CMAQ. Considering the medium computing resources and low model biases, the ML method is recommended for weather normalization of PM$_{2.5}$. Sensitivity analysis further suggested that the ML model with optimized hyperparameters and the exclusion of temporal variables in weather normalization can further produce reasonable results in emission-related trends of PM$_{2.5}$.

INTRODUCTION

Air pollution is a major environmental issue faced by heavily polluted regions around the world, including Central and South-Eastern Asia[1–3]. Reducing the number of premature deaths caused by air pollution has been identified as one of the United Nations' Sustainable Development Goals[4]. The new air quality guideline set by the World Health Organization has revised the annual concentration of fine particulate matter (PM$_{2.5}$) from 10 µg m^{-3} to 5 µg m^{-3} (ref.[5]), which requires further tightening of the measures for air pollution prevention and control[6]. Long-term observations of air pollutants capturing changes in air pollution can be used to evaluate the effectiveness of air quality policies[7–10]. The changes in air pollutant concentrations, however, are impacted both by emissions[11–13] and meteorological conditions[13–16]. Using the observed air pollutant concentrations without consideration of meteorological impacts to directly evaluate the effectiveness of measures has been questioned[17,18]. Therefore, assessing the effectiveness of air quality policies needs to decouple the impacts of emissions and meteorology on air pollutant variations.

Generally, there are three methods to estimate meteorology-normalized air pollutant variations (Supplementary Table 1). One is to use the chemistry transport models (CTMs) such as the Weather Research and Forecasting Model-Community Multiscale Air Quality

Model (WRF-CMAQ) and GEOS-Chem (GC). Zhang et al.[11] reported that the decrease in PM$_{2.5}$ in China was predominantly attributed to anthropogenic emissions abatement during 2013–2017 using WRF-CMAQ. With GEOS-Chem, Qiu et al.[19] quantified the emission-driven trends of PM$_{2.5}$ and found a substantial reduction of PM$_{2.5}$ concentration in eastern and central China from 2013 to 2017. Due to the inherent assumptions, parameterizations, and simplifications of processes in CTMs[20,21], and large uncertainties in emission inventories[22,23], CTM outputs are subject to large uncertainty[24]. One alternative method is the traditional statistical method (TSM), such as multiple linear regression (MLR) and Kolmogorov–Zurbenko (KZ) filter. The MLR is widely used to separate the contributions of emissions and meteorology to variations of PM$_{2.5}$[25–27] and ozone (O$_3$)[28–30]. The KZ filter developed by Rao and Zurbenko (1994) was first used to detect and track changes in O$_3$ in the US[31]. Since then, the KZ filter has been used in determining long-term trends of other air pollutants[32–34]. The other method is machine-learning (ML), a branch of statistical methods[35]. For instance, Grange et al.[36] developed a method for weather normalization of inhalable particulate matter (PM$_{10}$) by random forest (RF). Since then, the ML methods have been widely used especially during COVID-19 lockdowns[37–39]. Zheng et al.[37] found substantial reductions in air pollutant concentrations due to emission reductions during the

[1]Department of Atmospheric Sciences, School of Environmental Studies, China University of Geosciences (Wuhan), 430078 Wuhan, China. [2]Research Centre for Complex Air Pollution of Hubei Province, 430078 Wuhan, China. [3]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Key Laboratory for Aerosol-Cloud-Precipitation of China Meteorological Administration, PREMIC, Nanjing University of Information Science and Technology, 210044 Nanjing, China. [4]Earth and Environmental Sciences Programme and Graduate Division of Earth and Atmospheric Sciences, Faculty of Science, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR, China. [5]Anhui Institute of Meteorological Sciences, 230031 Hefei, China. [6]National Centre for Atmospheric Science (NCAS), The University of Manchester, Manchester M13 9PL, UK. [7]Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, UK. [8]School of Geography, Earth and Environment Sciences, University of Birmingham, Birmingham B15 2TT, UK. [9]Department of Environmental Sciences, Faculty of Meteorology, Environment, and Arid Land Agriculture, King Abdulaziz University, PO Box 80203, Jeddah, Saudi Arabia. ✉email: kongshaofei@cug.edu.cn; r.m.harrison@bham.ac.uk
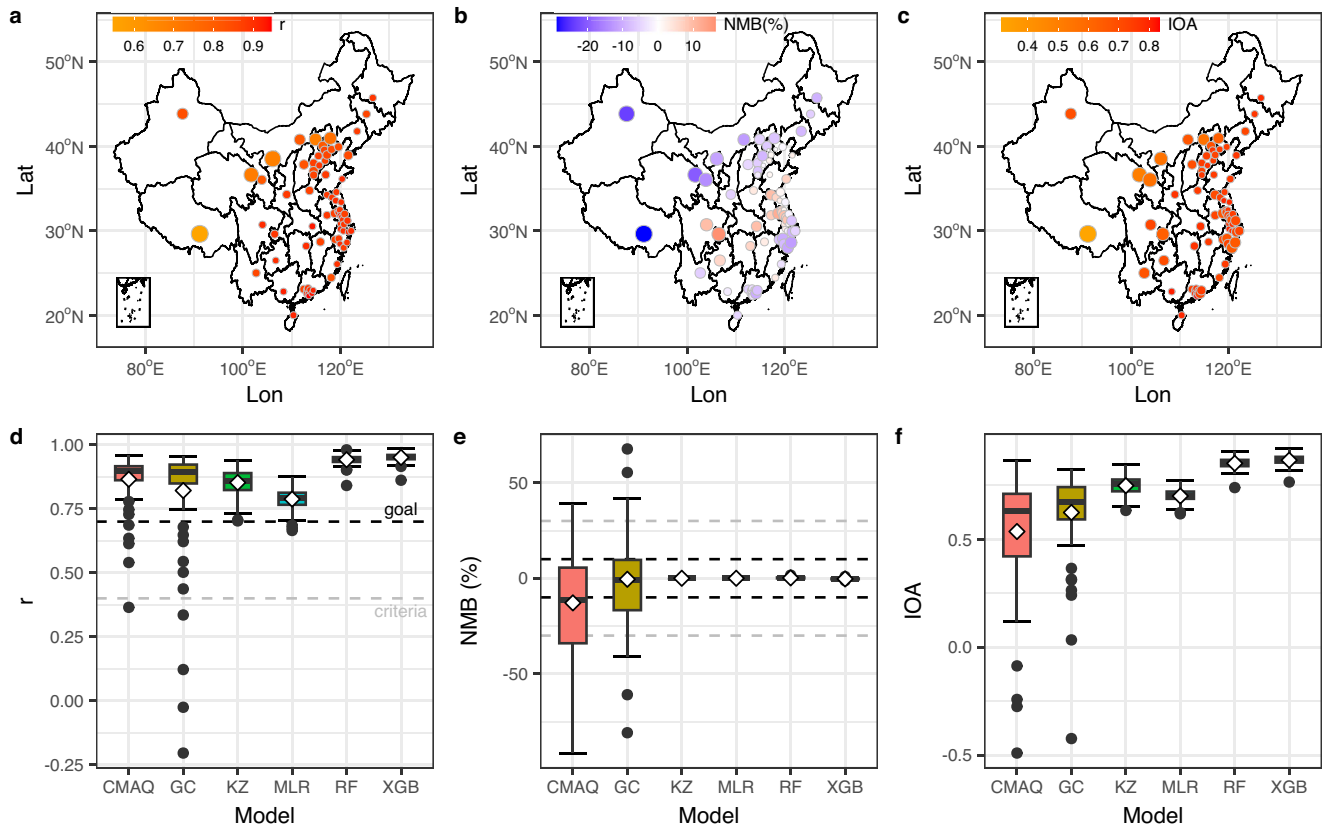
Fig. 1 Spatial distributions and boxplots of statistical metrics for model evaluation. a–c Spatial distributions in the average values of Pearson correlation coefficient (r), normalized mean bias (NMB), and index of agreement (IOA) derived from six methods. **d–f** Boxplots of r, NMB, and IOA for each method. The color and size of dots in the top panels are mapped to the mean values and standard deviations of statistics calculated from the six methods. The gray ("criteria") and black ("goal") horizontal dashed lines in the bottom panels represent the recommended benchmarks for model performance evaluation suggested by Emery et al.[48]. It should be noted that the hyperparameters for RF and XGB are tuned here for model evaluation.

lockdown period in Wuhan by the RF model. By the same method, Shi et al.[38] found abrupt but smaller-than-expected changes in surface air pollutant concentrations during COVID-19 in 11 cities globally. The ML methods are also used to assess the impacts of clean air actions on air pollutants. For instance, Vu et al.[40] used the RF model to assess the impacts of clean air action on air pollutant trends in Beijing between 2013 and 2017. Similarly, Dai et al.[41] answered the question of whether the Three-Year Action Plan improved the air quality in the Fenwei Plain of China by the RF model. Despite the wide adoption of traditional statistical and ML methods, the results from these two methods are always suspect due to their shortcomings in not considering the physical and chemical processes of air pollutants during their atmospheric lifetime.

Due to the high demands in running CTMs (e.g., air pollutant emission inventory, computer resources, and professional researchers), the application of CMTs is limited. As an alternative, TSM and ML have been widely used to normalize the weather on air pollutants. It should be noted that none of the existing methods is perfect in decoupling the impacts of emissions and meteorology on air pollutant variations[42]. The performance and comparability of different methods should therefore be assessed. Intra-comparisons between TSM[41–44] or intercomparisons between TSM, ML, and CTMs[40,45–47], however, are less reported[39,40]. One of the biggest challenges is the lack of simultaneous CTM results as a reference. The CTM simulations always focus on the study period's beginning and end year or specific months within each year, while the TSM and ML make use of the entire study period. Such differences in the study period

would introduce bias in intercomparison. If the performances and bias of different methods in decoupling the impacts of emissions and meteorology on air pollutant observations have been investigated, it will enhance our confidence to use these methods.

The notable air quality improvement in China from 2013 to 2017 has been acknowledged[11], which provides an opportunity to assess the performances of different methods in separating PM$_{2.5}$ variation drivers. The aims of this study are (1) assessing the differences in model performance of TSM, ML, and CTM methods in decoupling the impacts of meteorology and emissions on PM$_{2.5}$ and (2) comparing the trends (including emission-related and meteorology-related) of PM$_{2.5}$ and the bias of trends from statistical methods with the CTM result as a reference. The resources needed in different methods and three key factors that have impacts on weather normalization using the ML are discussed finally. This study would be beneficial to select a suitable method for investigating the long-term variations of aerosol compositions.

RESULTS

Performances of different models to reproduce PM$_{2.5}$ observation

Figure 1 shows the average values of statistical metrics between the observed and predicted PM$_{2.5}$ concentrations from different methods (the method-specific statistical metrics are shown in Supplementary Figs. 1–3). Overall, the metrics derived from the six methods were averaged (mean value ± standard deviation and hereafter) in the range of 0.55 ± 0.41 to 0.94 ± 0.04 for r,
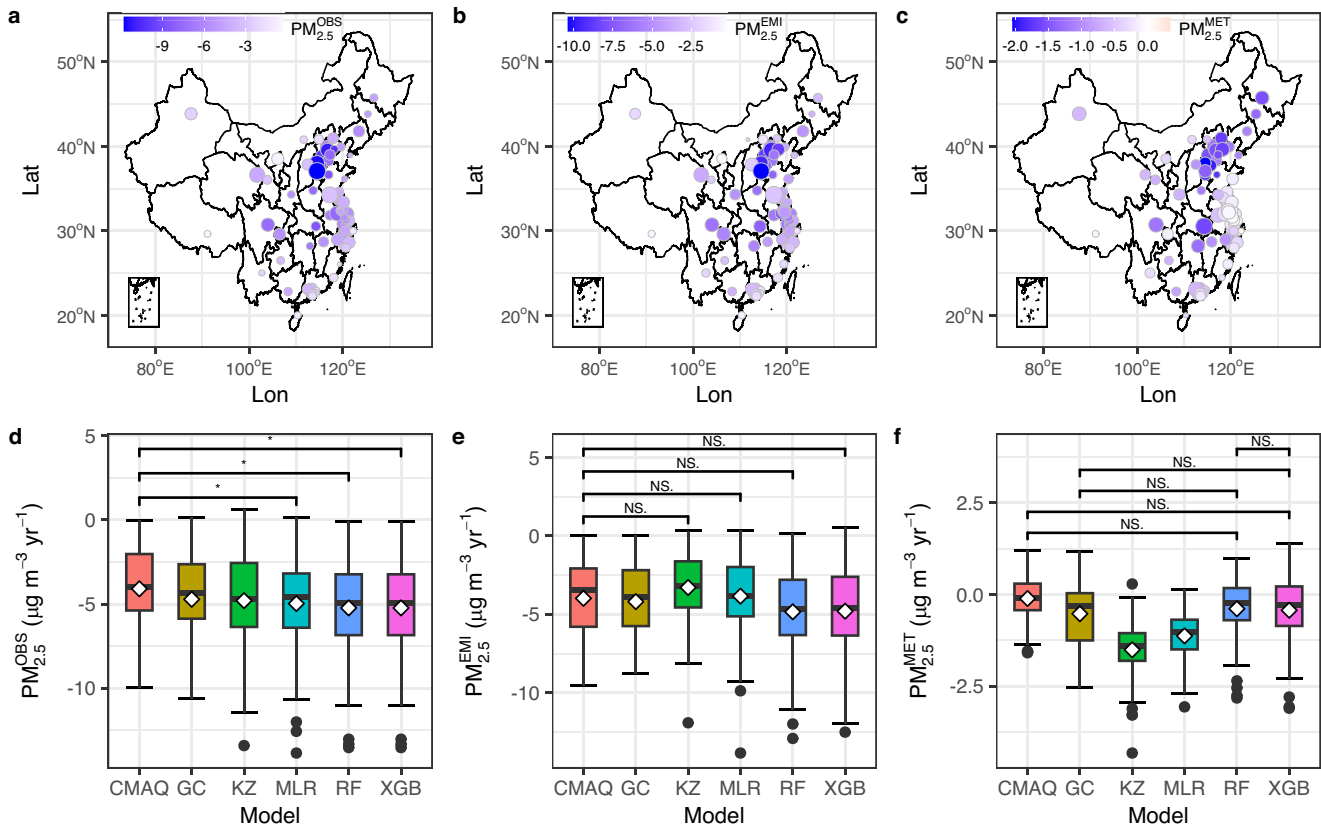
Fig. 2 Spatial distributions and boxplots of PM₂.₅ trends from different methods. a–c Spatial distributions in the average values of trends for PM₂.₅ observation ($PM_{2.5}^{OBS}$), emission-related PM₂.₅ ($PM_{2.5}^{EMI}$), and meteorology-related PM₂.₅ ($PM_{2.5}^{MET}$) derived from six methods. **d–f** Boxplots of $PM_{2.5}^{OBS}$, $PM_{2.5}^{EMI}$, and $PM_{2.5}^{MET}$ trends for each method. The meteorological conditions resampling strategy for the RF and XGB was from Grange et al.[36]. The color and size of dots in the top panels are mapped to the mean values and standard deviations of trends calculated from six methods. The marks in the bottom panels represent the differences between the two paired methods and the NS., and * mean the differences are not significant and significant at 0.05 levels.

16.2 ± 28.0% to 28.7 ± 44.6% for NMB, and 0.31 ± 0.60 to 0.83 ± 0.07 for index of agreement (IOA), respectively. It should be noted that the temporal resolution of data to calculate the statistics in Fig. 1 was monthly for CMAQ and GC, daily for MLR and KZ, and hourly for RF and extreme gradient boosting (XGB). If the temporal resolution of the data to calculate the statistical metrics for KZ, MLR, RF, and XGB was scaled to monthly, the TSM and ML showed even better performance to reproduce the observations (Supplementary Fig. 4). For instance, r values produced by MLR and RF significantly increased from 0.79 ± 0.04 to 0.85 ± 0.04 and 0.94 ± 0.02 to 1.0 ± 0.01, respectively, at the 0.001 level.

According to the "criteria" value of r greater than 0.4 and the "goal" value of NMB within ±30% for 24-h averaged PM₂.₅ evaluation[48], the MLR and KZ methods achieved acceptable performance in all cities. Most cities (71 of 74 sites for MLR) were close to the "goal" of evaluation with r greater than 0.7 and NMB within ±10% for statistical models. Similarly, the level of accuracy for RF and XGB models was considered to be close to the best a model can be expected to achieve. For that the r and NMB values calculated from hourly resolution data even fulfill the threshold of "goal" (Fig. 1d, e), not to mention the values calculated from daily data. r and NMB values for CMAQ and GC models calculated from monthly data meet the "criteria" of model evaluation for 47 and 63 cities, respectively. If the temporal resolution of data for CTM evaluation was changed to daily, the performances of CMAQ and GC would decline.

In terms of different methods to reproduce PM₂.₅ variations, the ML methods showed higher r and IOA values, with lower RMSE

values. The CTM, however, showed lower r and IOA values and higher RMSE compared to TSM and ML (Fig. 1d, f). For instance, the IOA values from different methods ranked as XGB (0.97 ± 0.01) > RF (0.96 ± 0.01) > KZ (0.76 ± 0.04) > MLR (0.74 ± 0.04) > GC (0.62 ± 0.19) > CMAQ (0.54 ± 0.26). A literature review in Supplementary Table 1 also showed a better performance of TSM and ML than CTM in reproducing the air pollutant concentrations. For instance, the correlation coefficient of the linear regression between the monthly observations and simulations of PM₂.₅ showed a higher value for the RF ($r^2 = 0.99$) model than CMAQ ($r^2 = 0.82$) in Beijing[40]. These statistical metrics for MLR, KZ, RF, and XGB models indicated that TSM and ML can capture the spatial-temporal variations of PM₂.₅ in this study.

Comparison in trends of PM₂.₅ from different methods

Supplementary Fig. 5 shows the time series of scaled PM₂.₅ concentrations derived from the six methods. Generally, all of the 74 cities showed a decreasing trend of PM₂.₅ with contributions from both emission-related and meteorology-related trends (Fig. 2). The trends of 74 cities were averaged as −11.8 ± 2.69 μg m⁻³ yr⁻¹ ~ −0.37 ± 0.36 μg m⁻³ yr⁻¹ for $PM_{2.5}^{OBS}$, −10.3 ± 2.66 μg m⁻³ yr⁻¹ ~ −0.27 ± 0.93 μg m⁻³ yr⁻¹ for $PM_{2.5}^{EMI}$, and −2.03 ± 0.80 μg m⁻³ yr⁻¹ ~ 0.33 ± 1.20 μg m⁻³ yr⁻¹ for $PM_{2.5}^{MET}$, respectively, from six methods. The high standard deviation of trends suggested the spatial heterogeneity in PM₂.₅ reduction during 2013–2017 in China (Supplementary Table 2). The high standard deviation of mean trends for PM₂.₅ in Fig. 2a–c was also related to the model differences (Supplementary Fig. 6). $PM_{2.5}^{OBS}$ calculated from CTMs had an

insignificant ($p > 0.05$) difference between CMAQ and GC. Similarly, $PM_{2.5}^{OBS}$ calculated from TSM (e.g., KZ and MLR) showed no statistical difference, and the same result for the ML (RF vs. XGB) (Supplementary Fig. 7). The trends of $PM_{2.5}^{OBS}$ from CTM, TSM, and ML, however, showed significant differences. The trend of $PM_{2.5}^{OBS}$ derived from CMAQ ($-4.09 \pm 2.44 \, \mu g \, m^{-3} \, yr^{-1}$) was significantly higher (less negative) than MLR ($-4.97 \pm 2.87 \, \mu g \, m^{-3} \, yr^{-1}$), RF ($-5.23 \pm 2.96 \, \mu g \, m^{-3} \, yr^{-1}$), and XGB ($-5.23 \pm 2.96 \, \mu g \, m^{-3} \, yr^{-1}$) at the 0.05 level (Fig. 2d). For $PM_{2.5}^{EMI}$ (Fig. 2e), the intra-comparison of trends within CTM, TSM, and ML showed no differences ($p > 0.05$). Intercomparison of $PM_{2.5}^{EMI}$ also showed insignificant ($p > 0.05$) differences between CMAQ ($- 3.98 \pm 2.19 \, \mu g \, m^{-3} \, yr^{-1}$), KZ ($-3.29 \pm 2.30 \, \mu g \, m^{-3} \, yr^{-1}$), MLR ($-3.84 \pm 2.54 \, \mu g \, m^{-3} \, yr^{-1}$), RF ($-4.84 \pm 2.79 \, \mu g \, m^{-3} \, yr^{-1}$), and XGB ($-4.80 \pm 2.78 \, \mu g \, m^{-3} \, yr^{-1}$). For $PM_{2.5}^{MET}$ (Fig. 2f), trends from CTM and ML showed insignificant differences while the trends from TSM were significantly lower than the other methods. No significant differences between the trends in $PM_{2.5}^{EMI}$ from TSM, ML, and CMAQ models suggesting the lack of physical-chemical mechanisms was not important at least in revealing the emission-related trends of $PM_{2.5}$ on the national scale by the TSM and ML.

Contributions of emission and meteorology to $PM_{2.5}$ trend by different methods

Using the scatterplot between $PM_{2.5}^{EMI}$, $PM_{2.5}^{MET}$, and $PM_{2.5}^{OBS}$, the relative contributions of emissions and meteorology to the variations of $PM_{2.5}$ were quantified (Supplementary Fig. 8). A contribution of $PM_{2.5}^{EMI}$ to $PM_{2.5}^{OBS}$ less than 100% indicates that the inter-annual variations of meteorology contribute to the reduction of $PM_{2.5}$. On the contrary, a percentage of $PM_{2.5}^{EMI}$ greater than 100% suggests the inter-annual variations of meteorology offset the reduction of $PM_{2.5}$ from emission variations. On the national scale, the decrease in $PM_{2.5}$ from 2013 to 2017 in China was dominated by emission reductions with contributions of 78.9% (KZ) ~90.5% (RF) according to the six modeling results (Supplementary Table 3). The comparable results between TSM, ML, and CTM suggested their ability to determine the dominant factor to variations of $PM_{2.5}$ at a large spatial scale.

The estimated contributions of emissions and meteorology to variations in $PM_{2.5}$ by different methods, however, showed a regional difference (Supplementary Table 3). For instance, the relative contributions of $PM_{2.5}^{EMI}$ to $PM_{2.5}^{OBS}$ calculated from CTM and ML were higher than 100% in YRD, suggesting the negative role of meteorology on $PM_{2.5}$ reduction from 2013 to 2017 (Supplementary Fig. 6). The percentages of $PM_{2.5}^{MET}$ to $PM_{2.5}^{OBS}$ from TSM, however, suggested the meteorology variations contributed to the reduction of the observed $PM_{2.5}$ in the YRD region, similar with previous studies with CTM adopted in YRD[49,50]. The opposite role of meteorology to the variation of $PM_{2.5}$ derived from different methods here and previous studies suggested that none of the existing methods can perfectly decouple the effects of emissions and meteorology on the trends of air pollutant concentrations. The different methods demonstrated comparable results in quantifying the influence of meteorological factors on $PM_{2.5}$ variations at the national scale, whereas differences were observed at a regional scale. Therefore, results from multiple methods (linear/non-linear) should be cross-checked to carefully evaluate the impacts of policies or interventions on regional air pollutant concentrations.

Bias in trends of $PM_{2.5}$ from different methods compared to CMAQ

With an assumption that the emission constant sensitivity simulation of a CTM (e.g., CMAQ in this study) produced a conceptual minimum of estimation error[19], the biases in trends (defined as $100\% \times (1 - $ the slope of a linear regression between CMAQ and other methods)) from the other five methods relative to CMAQ were calculated (Supplementary Fig. 9). $PM_{2.5}^{OBS}$ trends calculated from KZ and MLR were underestimated by 7% and 3%, while the trends from the other three methods were unbiased. Compared to $PM_{2.5}^{EMI}$ from CMAQ, trends from the other five methods showed underestimation with KZ underestimated most by 23%, followed by MLR (13%), XGB (3%), RF (2.8%), and GC (2.4%). Trends of meteorology-related $PM_{2.5}$ calculated from statistical and machine learning methods were highly biased with underestimation of 79%, 66%, 30%, and 28% for KZ, MLR, RF, and XGB, respectively. The bias of $PM_{2.5}^{MET}$ trend from GC, however, was overestimated by 6%.

The higher biases in trends of $PM_{2.5}$ from TSM were related to the model performance to reproduce the relationship between $PM_{2.5}$ and meteorological variables. Specifically, $PM_{2.5}^{EMI}$ was calculated from the residuals of linear fitting models for KZ and MLR methods (see Supplementary Methods for details), the higher residual or lower slope of the fitting in KZ and MLR methods was, the lower bias in $PM_{2.5}^{EMI}$ yielded (Supplementary Fig. 10a, b). The higher model performance of the KZ filter compared to MLR (e.g., r values of 0.85 ± 0.05 for the KZ filter vs. 0.79 ± 0.04 for MLR) in reproducing the relationship between meteorological variables and $PM_{2.5}$ can explain the larger bias in emission-related trend of $PM_{2.5}$ from KZ filter. A sensitivity study using the RF instead of the MLR model to build the relationship between the baseline component of $PM_{2.5}$ and meteorological factors in the KZ method further indicated a higher bias from fitting by the RF model, which showed higher slope and low residuals (Supplementary Fig. 10c). The lower biases in $PM_{2.5}$ trends from ML methods were possibly related to the inclusion of temporal variables (proxies of emission) in model training. The sensitivity analysis of bias in $PM_{2.5}^{EMI}$ from the RF model showed a larger bias from the model without the temporal variables included in model training (Supplementary Fig. 10d).

The biases of different methods were also related to the inherent uncertainties of CMAQ, which originated from uncertainties in air pollutant emission inventories and incomplete physical-chemical mechanisms. For instance, using the results from TSM and ML methods as the references, the higher biases of $PM_{2.5}^{EMI}$ and $PM_{2.5}^{MET}$ were produced for CTM (Supplementary Fig. 11). Additionally, these statistical methods assume that the variation of air pollutant is a linear sum of meteorological and emission changes[18,42]. Therefore, the influence of meteorological and emission changes on air pollutants can be cleanly separated from each other[19]. The impacts of meteorological variation may not be distinguishable from air pollutant trends driven by emission changes, due to their interactions[19]. Nevertheless, the ML methods perform robustly in getting the weather-normalized trends of $PM_{2.5}$ compared to TSM. Similar findings were also reported from previous studies, e.g., the widely used MLR did not perform well in correcting for emission-related and meteorology-related trends of air pollutants[19].

DISCUSSION

The required input datasets, advantages, disadvantages, biases in trends, and scopes of applications for different methods are summarized in Table 1. Compared to CTMs, the superiority of TSM and ML in weather normalization of air pollutants is less in required input datasets and their higher running speed with fewer computational resources (see Supplementary Note 1 and Supplementary Fig. 12 for details). The fewer required sources for running TSM and ML means that these methods can be run on personal computers, indicating their wide potential applications. Although TSM and ML have disadvantages in considering physical-chemical processes in their applications, these limitations are not significant in capturing the trend of $PM_{2.5}$ as shown above. Among the TSM and ML, the TSM has to address assumptions such as sample normality, homoscedasticity, independence, strict

Table 1. Summary of different models in weather normalization of $PM_{2.5}$.

Model	Required datasets	Advantages	Disadvantages	Bias in trend	Scopes of application
CTM	Emission inventories. Meteorology fields. Initial conditions. Boundary conditions.	4-D simulation results. Physical-chemical mechanisms encoded.	Heavy computation resources. Slow in computation. Accuracy highly depends on the accuracies of inputs. Computer cluster needed.	a	Large spatial scale simulation. Relative shorter study period depends on available resources.
MLR	Air pollutant observation. Meteorology variables.	Fast in computation. Fewer computation resources.	Address assumptions required for statistical models. Lack of physic-chemical mechanisms.	$PM_{2.5}^{EMI}$ (13%) $PM_{2.5}^{MET}$ (66%)	Site (city) or grid scales. Relative longer study period depends on observation data.
KZ-MLR	Air pollutant observation. Meteorology variables.	Fast in computation. Fewer computation resources.	Address assumptions required for statistical models. Lack of physical-chemical mechanisms.	$PM_{2.5}^{EMI}$ (23%) $PM_{2.5}^{MET}$ (79%)	Site (city) or grid scales. Relative longer study period depends on observation data.
RF	Air pollutant observation. Meteorology variables.	Fast in computation. Medium computation resources.	Lack of physical-chemical mechanisms. Performances depend on hyperparameters.	$PM_{2.5}^{EMI}$ (3%) $PM_{2.5}^{MET}$ (29%)	Site (city) or grid scales. Relative longer study period depends on observation data.
XGB	Air pollutant observation. Meteorology variables.	Fast in computation. Medium computation resources.	Lack of physical-chemical mechanisms. Performances depend on hyperparameters.	$PM_{2.5}^{EMI}$ (3%) $PM_{2.5}^{MET}$ (28%)	Site (city) or grid scales. Relative longer study period depends on observation data.

aUsed as a reference to evaluate other methods.

adherence to parametric requirements, and interaction effects among variables[51]. The ML is non-parametric and has the critical advantage of not needing to address many of the assumptions required for statistical methods[36]. Considering the application conditions, the balance between model performance and required resources, and their biases in normalizing the impacts of weather on $PM_{2.5}$, machine learning methods are recommended.

To better apply ML methods in the weather normalization of air pollutants, three influencing factors have been discussed in this study. Parameter setting is crucial in the ML methods to achieve optimal learning capacity during the training process and to achieve the best prediction performance during the testing stage[39,52]. The reported papers usually adopt the fixed parameters for the RF model training[10,36–38,40,53]. As shown in Fig. 3a, r and IOA calculated from the RF model with parameters tuned significantly increased compared to the RF model without parameters tuned. Trends of $PM_{2.5}^{OBS}$ and $PM_{2.5}^{EMI}$ from the tuned and untuned RF models showed insignificant differences, while the trends of $PM_{2.5}^{MET}$ showed a significant difference with a higher reduction rate from the tuned RF model (Fig. 3b). Compared to the results from CMAQ, the tuned RF model can reduce the bias of $PM_{2.5}^{EMI}$ trend by 9% from 12% to 3% (Fig. 3c). The bias in meteorology-related $PM_{2.5}$ trend for the tuned RF model also reduced by 12% from 41% to 29% compared to the untuned RF model (Supplementary Fig. 13). The bias in RF model with GC as a reference also verified the improvement of weather normalization of $PM_{2.5}$ by the tuned RF model (Supplementary Fig. 13). Therefore, the parameters for the ML methods are recommended to be optimized before application.

The meteorology resampling strategies adopted by the ML also influence the weather normalization result[54]. The widely reported meteorology resampling strategies include the method developed by Grange et al.[36] and Vu et al.[40] (denoted as "G" and "V", respectively, and hereafter). These two strategies have a shortage of comparing the trends from the CTMs. The meteorological factors in a CTM sensitivity simulation are fixed at a specific year while the meteorological variables in methods G and V are randomly sampled from the entire study period. We developed a resampling strategy (denoted as the "M" method hereafter, see Supplementary Note 2 for details), which resampled the meteorological variables from the year that was used for the sensitive simulation for CTM, e.g., 2017 in this study. As shown in Fig. 4a, changes in $PM_{2.5}^{EMI}$ trends (calculated as $PM_{2.5}^{EMI}$ trends from different resampling strategies − $PM_{2.5}^{EMI}$ trend from CMAQ) showed insignificant differences among different strategies. The biases in $PM_{2.5}^{EMI}$ trends with different strategies were underestimated by 2.22% (V30) ~ 18.5% (M) compared to the CMAQ reference result. The insignificant differences and low biases of $PM_{2.5}^{EMI}$ trends from different resampling strategies compared to CMAQ indicated these strategies can both produce reasonable emission-related trends of $PM_{2.5}$. Unlike the insensitivity of emission-related trends to different resampling strategies, the meteorology-related $PM_{2.5}$ trends were more sensitive to resampling strategies (Fig. 4b). For instance, trends of $PM_{2.5}^{MET}$ with G, V5, and V30 strategies were lower than those from CMAQ with a bias of 27.7%, 40.2%, and 45.8% respectively. Given the fact that the insignificant differences, low bias of trends, easy and fast calculation properties of the meteorology resampling strategy developed by Grange et al.[36], this strategy is recommended for weather normalization of air pollutants.

The inclusion or exclusion of the temporal variables (e.g., Julian day and the day of the week in this study) in the prediction process should also be emphasized. Previous studies disagree with each other in the inclusion of temporal variables. Grange et al.[36] recommended the inclusion of temporal variables in weather normalization, and similar research was reported elsewhere[41,55,56]. In this strategy, meteorological variables and temporal variables were randomly sampled and they were used to predict the $PM_{2.5}$
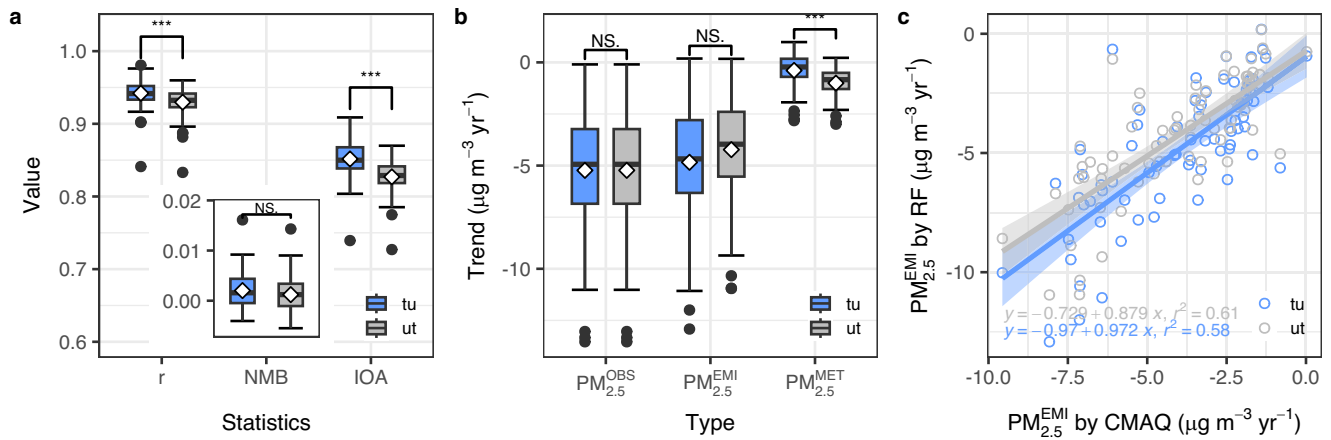
Fig. 3 Comparison in results derived from the random forest model with hyperparameters tuned and untuned. a Boxplots of statistics derived from the tuned RF model (tu) and untuned RF model (ut). **b** Boxplots of the trends for PM$_{2.5}$ observation (PM$_{2.5}^{OBS}$), emission-related PM$_{2.5}$ (PM$_{2.5}^{EMI}$), and meteorology-related PM$_{2.5}$ (PM$_{2.5}^{MET}$); **c** Scatterplot between emission-related trends of PM$_{2.5}$ derived from RF and CMAQ models. NS. and *** represent the differences between the two groups that are not significant ($p > 0.05$) and are significant at 0.001 levels.
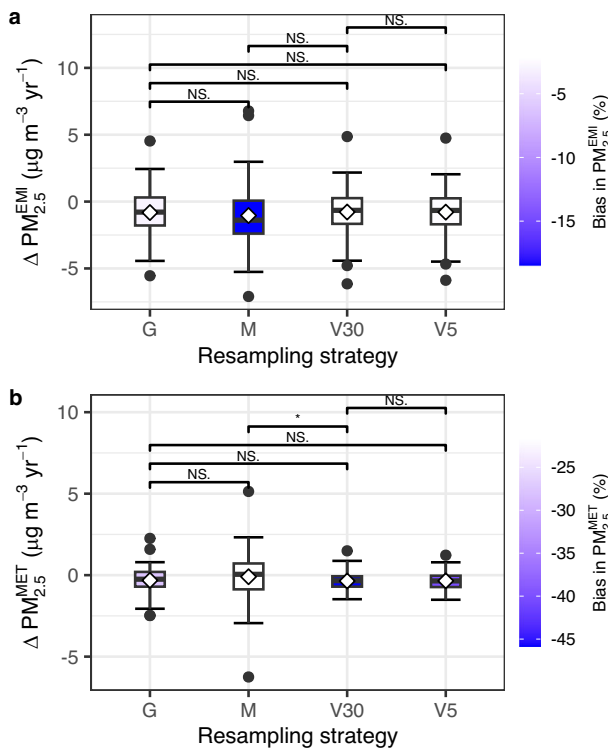

Fig. 4 Comparisons in changes of PM$_{2.5}$ trends from different meteorology resampling strategies with CMAQ as a reference. a Emission-related trend of PM$_{2.5}$. **b** Meteorology-related trend of PM$_{2.5}$. The filled colors in boxes represent the bias (%) calculated as 100% × (1 – the slope of a linear regression between the CMAQ and XGB model with different resampling strategies). NS. and * represent the differences between the two groups that are not significant ($p > 0.05$) and significant at 0.05 levels. The details about the resampling strategy can be found in Supplementary Note 2.

concentration. On the contrary, the exclusion of temporal variables (randomly sampled meteorological variables and fixed temporal variables) in weather normalization[40,54] was also adopted. Using the sensitive experiment with temporal variables included (RF_wt) and not included (RF_nt), the impact of this factor was discussed here. As shown in Fig. 5a, the variation of

PM$_{2.5}^{EMI}$ from RF_wt was continuously decreased, while the time series of PM$_{2.5}^{EMI}$ from RF_nt showed periodic decreases from 2013 to 2017. The linear regression between PM$_{2.5}^{EMI}$ trends from RF_wt and CMAQ showed a slope approaching 1, which was higher than the slope for fitting between RF_nt and CMAQ (Fig. 5b). This was due to the time series of PM$_{2.5}^{EMI}$ from RF_wt was more coincided with CMAQ compared to that from RF_nt. The temporal variables can be used as proxies for cyclical emission patterns[8] and if the temporal variables were randomly sampled in prediction, the signal of emission variations was erased from the normalized time series. As a result, the time series of RF_wt well revealed the long-term emissions[54] while the time series of RF_nt was able to characterize the seasonal and long-term emission trends[40,54]. In reality, air pollutant emissions have seasonal variations that arise from energy consumption patterns (e.g., heating during the cold season)[33,57]. Therefore, the results from the resampling strategy with temporal variables excluded were more reasonable despite it having a higher bias in trend of PM$_{2.5}^{EMI}$ with CMAQ as a reference.

Traditional statistical methods and machine learning have been widely used for weather normalization of primary and secondary air pollutants (Supplementary Table 1). These methods may result in different outcomes for the impacts of meteorology on air pollutants vary with the properties of air pollutants[19]. Particularly, separating and quantifying the effects of meteorology on O$_3$ is more challenging due to the complex interaction between meteorology, emissions, and chemical formation[58,59]. With CTM as a reference, the performance of TSM and ML in weather normalization of other air pollutants should be investigated before their application. With the successful reconstruction of air pollutant datasets derived from satellites[60–62], CTM[63,64], and ground observations[63–65], the long-term and full-coverage datasets are developed. Coupled with these open-accessed datasets (e.g., the China High Air Pollutants (CHAP), Track Air Pollution in China (TAP), and MERRA-2), ML method (e.g., RF and XGB), and recommendations in this study (e.g., hyperparameters tuning, meteorological condition resampling strategy, and exclusion of temporal variables in resampling), evaluating the effectiveness of air pollution prevention and climate change response policies can be conducted at regional and national scales. These potential evaluations contribute to solving air pollution and fulfilling the United Nations' Sustainable Development Goals.
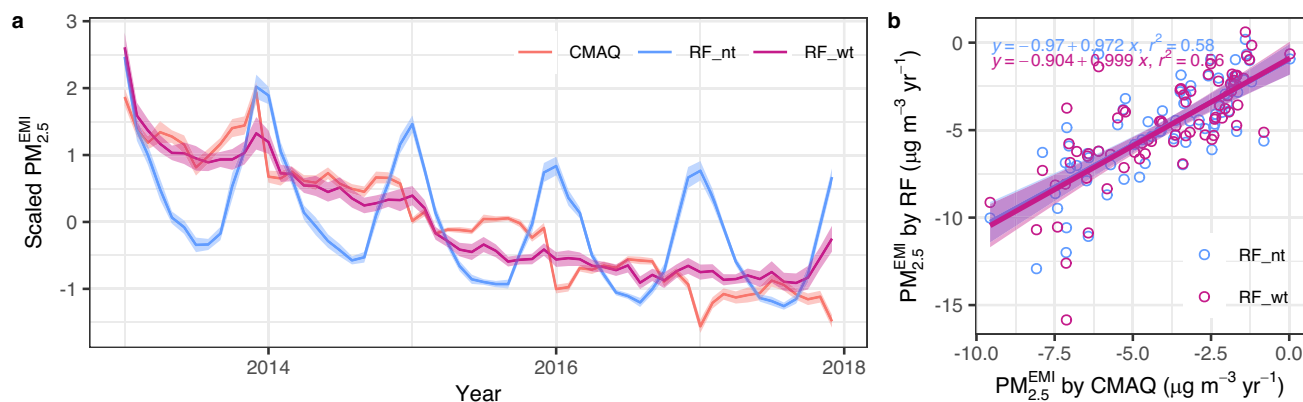
Fig. 5   Comparisons in results of random forest model with temporal variables included and excluded in weather normalization. a Time series of scaled monthly values in $PM_{2.5}^{EMI}$ from different methods. b Scatterplot between $PM_{2.5}^{EMI}$ trends from CMAQ and RF with temporal variables included (RF_wt) and excluded (RF_nt).

METHODS

Data sources and preprocessing

Hourly ground observations of $PM_{2.5}$ during 2013–2017 were from the national air quality monitoring network established and operated by China National Environmental Monitoring Center. 74 key cities (Supplementary Fig. 14) were selected in this study due to their data availability from 2013 to 2017. Data quality control was conducted according to previous studies[34,66,67] (see Supplementary Methods for details). Hourly values of meteorological variables including temperature at 2 m (T2M), dewpoint at 2 m (D2M), mean sea-level pressure (MSL), eastward and northward wind components of wind at 10 m (U10, V10), total precipitation (TP), boundary layer height (BLH), total cloud cover (TCC), and surface downward solar radiation (SSR) were obtained from the ERA-5 single-level pressure reanalysis datasets[68]. The relative humidity (RH) was calculated with T2M and D2M[69]. The monthly mean concentrations of $PM_{2.5}$ from 2013 to 2017 by WRF-CMAQ and GEOS-Chem were from Zhang et al.[11] and Zhai et al.[70], respectively. More details about the input meteorology, emission inventory, and simulation settings (base + sensitivity) by CTM can be found in the references above. The methods for calculating the emission-related and meteorology-related $PM_{2.5}$ concentrations from CTMs can be found in the Supplementary Methods.

Weather normalization of $PM_{2.5}$ by TSM and ML

To decouple the effects of meteorology on $PM_{2.5}$ variations, two traditional statistical methods (MLR, KZ) and machine learning methods (RF and XGB) were adopted using the meteorological variables mentioned above in each city. Specifically, the T2M, MSL, U10, V10, RH, TP, BLH, TCC, and SSR were used to build the MLR and KZ filter models. In addition to these meteorological variables, time variables (Unix time: number of seconds since 1970-1-1, Julian day: day of the year, day of the week) acted as emission proxies[8], and clusters of backward trajectories reaching each city acted as transport indicator[36,38] were also used in RF and XGB models. For MLR and KZ model building, the daily averages of air pollutants and meteorological variables were used, while the hourly observations were used in RF and XGB models. A flow chart to show the weather normalization of $PM_{2.5}$ using different methods is shown in Supplementary Fig. 15. The data process, model building, and weather normalization are described below in detail.

MLR.   Following previous studies[26,28,30] but with a little modification, nine meteorological variables mentioned above were used to establish the relationship between meteorological factors and $PM_{2.5}$, instead of employing a stepwise MLR to exclude less

important variables. The anomalies of meteorological conditions and $PM_{2.5}$ were obtained by moving the 5-year mean values of 50-d moving averages from the 10-d mean time series and the anomalies calculated by this method were deseasonalized but not detrended. According to a previous study[26], the 50-d moving window was chosen here because the anomalies of $PM_{2.5}$ and other meteorological variables calculated in this manner were not sensitive to the moving window (Supplementary Fig. 16). The anomalies of $PM_{2.5}$ and meteorological variables were finally used to build the MLR model. The prediction of MLR was considered as the meteorology-driven $PM_{2.5}$ concentration and the residuals of fitting were considered as the $PM_{2.5}$ concentration attributed to emission changes[26,30]. More details about MLR to separate the meteorology and emission-related $PM_{2.5}$ concentrations can be found in Supplementary Methods.

KZ.   The KZ filter ($KZ_{(m, p)}$) uses different iteration times (p) and moving averages of time width (m) to separate the time series of air pollutant into different components, e.g., $KZ_{(365, 3)}$ to filter out long-term component[31,71,72], and $KZ_{(15, 5)}$ to get the baseline component (seasonal + long-term components)[33]. To get the long-term component of $PM_{2.5}$ and its two subcomponents including emission-related and meteorology-related, the baseline components of $PM_{2.5}$ and meteorological factors were used to build the MLR model with $PM_{2.5}$ as the dependent variable. The emission-related concentration was obtained by $KZ_{(365, 3)}$ to the residuals of MLR above. The meteorology-related concentration was calculated as the difference between the long-term concentration of $PM_{2.5}$ and the emission-related concentration. More details about KZ-MLR can be found in Supplementary Methods.

RF.   The meteorological variables, time variables, and cluster of trajectories mentioned above were used to build the RF model. Before model training, the dataset was randomly divided into two sub-datasets with a ratio of 7:3. 70% of the datasets were used to build the model and the remaining 30% of datasets were used to test the model. In line with previous studies[36,38,53,73,74], the settings below were used to train the RF model: the number of the tree (ntree) = 300; the number of variables that may split at each node (mtry) = 3; the minimum size of terminal nodes (min.node.size) = 5. In addition to the default settings for the RF model, these parameters were also tuned by random search with 5-fold cross-validation after 100 times evaluation. The 5-fold cross-validation was used here to determine the optimal hyperparameter combinations[75,76]. The search space consisted of ntree, mtry, and min.node.size with their ranges of 10 ~ 1000, 1–13, and 1–13, respectively. The results of tuned hyperparameters for the RF model are provided in Supplementary Table 4. After the model

Table 2. Experiment design and its purpose in this study.

Experiment	Method	Tuned[b]	Experiment design	Purpose
1#	CMAQ[a]			Intercomparison of different methods
	GC			
	KZ			
	MLR			
	RF	Yes	G[c]	
	XGB	Yes	G	
2#	RF	No	G	The effect of hyperparameter tuning on weather normalization result
	RF	Yes	G	
3#	XGB	Yes	G	The effect of meteorology resampling strategy on weather normalization result
	XGB	Yes	V5[d]	
	XGB	Yes	V30[e]	
	XGB	Yes	M[f]	
4#	RF	Yes	G with temporal variables[g]	The effect of the temporal variables on weather normalization result
	RF	Yes	G without temporal variables	

[a] Used as a reference to evaluate the bias of other methods.
[b] Designed for machine learning.
[c–f]Different meteorology resampling strategies. V5 and V30 represent the meteorological variables that are randomly sampled from 2013–2017 and 1988–2017, respectively. More details can be found in Supplementary Note 2.
[g]Temporal variables include Julian's day and day of the week in this study.

training, the weather normalization in each observation was conducted by randomly sampling the meteorological variables from the meteorological data pool without replacement to predict the concentration by 500 times (sensitive analysis of resampling times on result is provided in Supplementary Methods and Supplementary Fig. 17). The weather-normalized concentration (emission-related concentration) for each observation was finally calculated as the arithmetic mean of 500 predictions. The meteorology-related $PM_{2.5}$ concentration was then calculated as the difference between observed air pollutant concentration and emission-related concentration[13,37,53,73].

XGB. Similar to weather normalization using RF, the XGB tree model was used to build the relation between hourly air pollutant concentrations and meteorological variables and other predictor variables[55]. Three key parameters including the number of gradient-boosted trees (nrounds), the maximum tree depth for base learners (max_depth), and boosting learning rate (eta)[77] were optimized by random search with 5-fold cross-validation[55,77]. The search space consisted of nrounds, max_depth, and eta with their ranges of 10 to 1000, 1 to 13, and 0 to 1, respectively. The terminator of the random search was chosen as 100 times evaluation. Based on the performance of the 5-fold cross-validation, the optimal hyperparameters were obtained in each city (Supplementary Table 4). After tuning, these parameters were used to train the XGB model. The trained XGB model was further used for weather normalization with the same process as described above in the RF model.

Experiment design. A series of calculations using different methods mentioned above were conducted to compare the performance of different methods in weather normalization of $PM_{2.5}$ (1# experiment in Table 2). To exclude the effects of the data split process on machine learning, the same data was used in model training and testing for RF and XGB models, e.g., the effects of hyperparameter tuning on the RF model (2# experiment). Additionally, the same trained model (e.g., XGB) was used in discussing the meteorology resampling strategy on weather normalization results (3# experiment). Finally, the 4# experiment

was designed to discuss the inclusion and exclusion of temporal variables on weather normalization results by the RF model.

Trend calculation and statistical parameters
Using the methods mentioned above, $PM_{2.5}$ observation ($PM_{2.5}^{OBS}$) was decoupled into emission-related ($PM_{2.5}^{EMI}$) and meteorology-related ($PM_{2.5}^{MET}$) concentrations. To make sure that the trends of $PM_{2.5}$ observation equaled the trends of $PM_{2.5}^{EMI}$ and $PM_{2.5}^{MET}$ from 2013 to 2017, the trends were calculated by linear regression between annual values in $PM_{2.5}$ concentrations and years[33,34,78]. The slope of the linear regression equation was regarded as the trend. To evaluate the model performances of different models to reproduce the observations of $PM_{2.5}$, several statistical parameters including the Pearson correlation coefficient (r), normalized mean bias (NMB), and IOA were used (see Supplementary Table 5 for more details).

REFERENCES
1. Cheng, Z. et al. Status and characteristics of ambient $PM_{2.5}$ pollution in global megacities. *Environ. Int.* **89–90**, 212–221 (2016).

2. Shaddick, G., Thomas, M. L., Mudu, P., Ruggeri, G. & Gumy, S. Half the world's population are exposed to increasing air pollution. *npj Clim. Atmos. Sci.* **3**, 23 (2020).

3. Sicard, P. et al. Trends in urban air pollution over the last two decades: a global perspective. *Sci. Total Environ.* **858**, 160064 (2023).

4. World Health Organization. *Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease*. (World Health Organization, 2016).

5. World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter (PM$_{2.5}$ and PM$_{10}$), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide: Executive Summary*. (World Health Organization, 2021).

6. Cheng, J. et al. Pathways of China's PM$_{2.5}$ air quality 2015–2060 in the context of carbon neutrality. *Natl. Sci. Rev.* **8**, nwab078 (2021).

7. Sullivan, T. J. et al. Air pollution success stories in the United States: The value of long-term observations. *Environ. Sci. Policy* **84**, 69–73 (2018).

8. Grange, S. K. & Carslaw, D. C. Using meteorological normalisation to detect interventions in air quality time series. *Sci. Total Environ.* **653**, 578–588 (2019).

9. Dai, Q. et al. Trends of source apportioned PM$_{2.5}$ in Tianjin over 2013–2019: impacts of clean air actions. *Environ. Pollut.* **325**, 121344 (2023).

10. Song, C. et al. Attribution of air quality benefits to clean winter heating policies in China: combining machine learning with causal inference. *Environ. Sci. Technol.* **57**, 17707–17717 (2023).

11. Zhang, Q. et al. Drivers of improved PM$_{2.5}$ air quality in China from 2013 to 2017. *Proc. Natl. Acad. Sci. USA* **116**, 24463–24469 (2019).

12. Joshi, R. et al. Direct measurements of black carbon fluxes in central Beijing using the eddy covariance method. *Atmos. Chem. Phys.* **21**, 147–162 (2021).

13. Zheng, H. et al. Enhanced ozone pollution in the summer of 2022 in China: the roles of meteorology and emission variations. *Atmos. Environ.* **301**, 119701 (2023).

14. Wang, T. et al. Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects. *Sci. Total Environ.* **575**, 1582–1596 (2017).

15. Chen, Z. et al. Understanding meteorological influences on PM$_{2.5}$ concentrations across China: a temporal and spatial perspective. *Atmos. Chem. Phys.* **18**, 5343–5358 (2018).

16. Chen, Z. et al. Influence of meteorological conditions on PM$_{2.5}$ concentrations across China: a review of methodology and mechanism. *Environ. Int.* **139**, 105558 (2020).

17. Zhong, Q. et al. Distinguishing emission-associated ambient Air PM$_{2.5}$ concentrations and meteorological factor-induced fluctuations. *Environ. Sci. Technol.* **52**, 10416–10425 (2018).

18. Zhong, Q. et al. PM$_{2.5}$ reductions in Chinese cities from 2013 to 2019 remain significant despite the inflating effects of meteorological conditions. *One Earth* **4**, 448–458 (2021).

19. Qiu, M., Zigler, C. & Selin, N. E. Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. *Atmos. Chem. Phys.* **22**, 10511–10566 (2022).

20. Manders, A. M. M. et al. The impact of differences in large-scale circulation output from climate models on the regional modeling of ozone and PM. *Atmos. Chem. Phys.* **12**, 9441–9458 (2012).

21. Otero, N. et al. A multi-model comparison of meteorological drivers of surface ozone over Europe. *Atmos. Chem. Phys.* **18**, 12269–12288 (2018).

22. Zhao, Y., Nielsen, C. P., Lei, Y., McElroy, M. B. & Hao, J. Quantifying the uncertainties of a bottom-up emission inventory of anthropogenic atmospheric pollutants in China. *Atmos. Chem. Phys.* **11**, 2295–2308 (2011).

23. Li, M. et al. Anthropogenic emission inventories in China: a review. *Natl. Sci. Rev.* **4**, 834–866 (2017).

24. Sokhi, R. S. et al. Advances in air quality research–current and emerging challenges. *Atmos. Chem. Phys.* **22**, 4615–4703 (2022).

25. Tai, A. P. K., Mickley, L. J. & Jacob, D. J. Correlations between fine particulate matter (PM$_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of PM$_{2.5}$ to climate change. *Atmos. Environ.* **44**, 3976–3984 (2010).

26. Zhai, S. et al. Fine particulate matter (PM$_{2.5}$) trends in China, 2013–2018: separating contributions from anthropogenic emissions and meteorology. *Atmos. Chem. Phys.* **19**, 11031–11041 (2019).

27. Chen, L., Zhu, J., Liao, H., Yang, Y. & Yue, X. Meteorological influences on PM$_{2.5}$ and O$_3$ trends and associated health burden since China's clean air actions. *Sci. Total Environ.* **744**, 140837 (2020).

28. Li, K. et al. Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China. *Proc. Natl. Acad. Sci. USA* **116**, 422–427 (2019).

29. Li, K. et al. A two-pollutant strategy for improving ozone and particulate air quality in China. *Nat. Geosci.* **12**, 906–910 (2019).

30. Li, K. et al. Increases in surface ozone pollution in China from 2013 to 2019: anthropogenic and meteorological influences. *Atmos. Chem. Phys.* **20**, 11423–11433 (2020).

31. Rao, S. T. & Zurbenko, I. G. Detecting and tracking changes in ozone air quality. *J. Air. Waste. Manage.* **44**, 1089–1092 (1994).

32. Henneman, L. R. F., Holmes, H. A., Mulholland, J. A. & Russell, A. G. Meteorological detrending of primary and secondary pollutant concentrations: method application and evaluation using long-term (2000–2012) data in Atlanta. *Atmos. Environ.* **119**, 201–210 (2015).

33. Seo, J. et al. Effects of meteorology and emissions on urban air quality: a quantitative statistical approach to long-term records (1999–2016) in Seoul, South Korea. *Atmos. Chem. Phys.* **18**, 16121–16137 (2018).

34. Zheng, H. et al. A 5.5-year observations of black carbon aerosol at a megacity in Central China: levels, sources, and variation trends. *Atmos. Environ.* **232**, 117581 (2020).

35. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edn. (Springer, 2009).

36. Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E. & Hueglin, C. Random forest meteorological normalisation models for Swiss PM$_{10}$ trend analysis. *Atmos. Chem. Phys.* **18**, 6223–6239 (2018).

37. Zheng, H. et al. Significant changes in the chemical compositions and sources of PM$_{2.5}$ in Wuhan since the city lockdown as COVID-19. *Sci. Total Environ.* **739**, 140000 (2020).

38. Shi, Z. et al. Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns. *Sci. Adv.* **7**, eabd6696 (2021).

39. Wong, Y. J. et al. Quantification of COVID-19 impacts on NO$_2$ and O$_3$: Systematic model selection and hyperparameter optimization on AI-based meteorological-normalization methods. *Atmos. Environ.* **301**, 119677 (2023).

40. Vu, T. V. et al. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmos. Chem. Phys.* **19**, 11303–11314 (2019).

41. Dai, X. et al. Has the three-year action plan improved the air quality in the Fenwei Plain of China? Assessment based on a machine learning technique. *Atmos. Environ.* **286**, 119204 (2022).

42. Lin, Y. et al. Decoupling impacts of weather conditions on interannual variations in concentrations of criteria air pollutants in South China–constraining analysis uncertainties by using multiple analysis tools. *Atmos. Chem. Phys.* **22**, 16073–16090 (2022).

43. Weng, X., Forster, G. L. & Nowack, P. A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019. *Atmos. Chem. Phys.* **22**, 8385–8402 (2022).

44. Ji, Y. et al. Using machine learning to quantify drivers of aerosol pollution trend in China from 2015 to 2022. *Appl. Geochem.* **151**, 105614 (2023).

45. Sun, X. et al. Meteorology impact on a receptor region in the regional transport of air pollutants: observational study of recent emission reductions in central China. *Atmos. Chem. Phys.* **22**, 3579–3593 (2022).

46. Fang, C., Qiu, J., Li, J. & Wang, J. Analysis of the meteorological impact on PM$_{2.5}$ pollution in Changchun based on KZ filter and WRF-CMAQ. *Atmos. Environ.* **271**, 118924 (2022).

47. Chen, Z. et al. The control of anthropogenic emissions contributed to 80% of the decrease in PM$_{2.5}$ concentrations in Beijing from 2013 to 2017. *Atmos. Chem. Phys.* **19**, 13519–13533 (2019).

48. Emery, C. et al. Recommendations on statistics and benchmarks to assess photochemical model performance. *J. Air. Waste. Manage.* **67**, 582–598 (2017).

49. Guan, P. et al. Assessment of emission reduction and meteorological change in PM$_{2.5}$ and transport flux in typical cities cluster during 2013–2017. *Sustainability* **13**, 5685 (2021).

50. Zhen, J., Guan, P., Yang, R. & Zhai, M. Transport matrix of PM$_{2.5}$ in Beijing-Tianjin-Hebei and Yangtze river delta regions: assessing the contributions from emission reduction and meteorological conditions. *Atmos. Environ.* **304**, 119775 (2023).

51. Immitzer, M., Atzberger, C. & Koukal, T. Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data. *Rem. Sensing* **4**, 2661–2693 (2012).

52. Zhu, J. J., Yang, M. & Ren, Z. J. Machine learning in environmental research: common pitfalls and best practices. *Environ. Sci. Technol.* **57**, 17671–17689 (2023).

53. Hou, L. et al. Revealing drivers of haze pollution by explainable machine learning. *Environ. Sci. Technol. Lett.* **9**, 112–119 (2022).

54. Wu, Q. et al. Evaluation of NO$_x$ emissions before, during, and after the COVID-19 lockdowns in China: a comparison of meteorological normalization methods. *Atmos. Environ.* **278**, 119083 (2022).

55. Li, C., Zhu, Q., Jin, X. & Cohen, R. C. Elucidating contributions of anthropogenic volatile organic compounds and particulate matter to ozone trends over China. *Environ. Sci. Technol.* **56**, 12906–12916 (2022).

56. Wang, M. et al. Slower than expected reduction in annual PM$_{2.5}$ in Xi'an revealed by machine learning-based meteorological normalization. *Sci. Total Environ.* **841**, 156740 (2022).

57. Zhu, D. et al. Temporal and spatial trends of residential energy consumption and air pollutant emissions in China. *Appl. Energy.* **106**, 17–24 (2013).

58. Porter, W. C. & Heald, C. L. The mechanisms and meteorological drivers of the summertime ozone–temperature relationship. *Atmos. Chem. Phys.* **19**, 13367–13381 (2019).

59. Lu, X., Zhang, L. & Shen, L. Meteorology and climate influences on tropospheric ozone: a review of natural sources, chemistry, and transport patterns. *Curr. Pollut. Rep.* **5**, 238–260 (2019).

60. Xiao, Q., Chang, H. H., Geng, G. & Liu, Y. An ensemble machine-learning model to predict historical PM$_{2.5}$ concentrations in China from satellite data. *Environ. Sci. Technol.* **52**, 13260–13269 (2018).

61. Xiao, Q. et al. Evaluation of gap-filling approaches in satellite-based daily PM$_{2.5}$ prediction models. *Atmos. Environ.* **244**, 117921 (2021).

62. Wei, J. et al. Satellite-derived 1-km-Resolution PM$_1$ Concentrations from 2014 to 2018 across China. *Environ. Sci. Technol.* **53**, 13265–13274 (2019).

63. Xue, T. et al. Spatiotemporal continuous estimates of PM$_{2.5}$ concentrations in China, 2000–2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations. *Environ. Int.* **123**, 345–357 (2019).

64. Geng, G. et al. Tracking air pollution in China: near real-time PM$_{2.5}$ retrievals from multisource data fusion. *Environ. Sci. Technol.* **55**, 12106–12115 (2021).

65. Zhong, J. et al. Robust prediction of hourly PM$_{2.5}$ from meteorological data using LightGBM. *Natl. Sci. Rev.* **8**, nwaa307 (2021).

66. Barrero, M. A., Orza, J. A. G., Cabello, M. & Cantón, L. Categorisation of air quality monitoring stations by evaluation of PM$_{10}$ variability. *Sci. Total Environ.* **524–525**, 225–236 (2015).

67. Song, C. et al. Air pollution in China: status and spatiotemporal variations. *Environ. Pollut.* **227**, 334–347 (2017).

68. Hersbach, H. et al. ERA5 hourly data on pressure levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.bd0915c6 (2023).

69. Dutton, J. A. *The Ceaseless Wind: an Introduction to the Theory of Atmospheric Motion*. (McGraw-Hill, 1976).

70. Zhai, S. et al. Control of particulate nitrate air pollution in China. *Nat. Geosci.* **14**, 389–395 (2021).

71. Rao, S. T., Zalewsky, E. & Zurbenko, I. G. Determining temporal and spatial variations in ozone air quality. *J. Air. Waste Manage.* **45**, 57–61 (1995).

72. Rao, S. T. et al. Space and time scales in ambient ozone data. *Bull. Amer. Meteor. Soc.* **78**, 2153–2166 (1997).

73. Zhang, Y. et al. Significant changes in chemistry of fine particles in wintertime Beijing from 2007 to 2017: impact of clean air actions. *Environ. Sci. Technol.* **54**, 1344–1352 (2020).

74. Song, C. et al. Understanding sources and drivers of size-resolved aerosol in the high Arctic islands of Svalbard using a receptor model coupled with machine learning. *Environ. Sci. Technol.* **56**, 11189–11198 (2022).

75. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).

76. Zhong, S. et al. Machine learning: new ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **55**, 12741–12754 (2021).

77. Zheng, Z., Zhao, L. & Oleson, K. W. Large model structural uncertainty in global projections of urban heat waves. *Nat. Commun.* **12**, 3736 (2021).

78. Bae, M., Kim, B. U., Kim, H. C., Kim, J. & Kim, S. Role of emissions and meteorology in the recent PM$_{2.5}$ changes in China and South Korea from 2015 to 2018. *Environ. Pollut.* **270**, 116233 (2021).

AUTHOR CONTRIBUTIONS

H.Z., S.K., and R.M.H. conceived the study, wrote, and revised the manuscript. S.Z provided the Geos-Chem simulation results. X.S. processed data. All co-authors reviewed and commented on the paper.

COMPETING INTERESTS

R.M.H. is co-editor-in-chief. He was not involved in the journal's review of or decisions related to this manuscript. The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-023-00536-7.

Correspondence and requests for materials should be addressed to Shaofei Kong or Roy M. Harrison.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.