Diagnosing drivers of $PM_{2.5}$ simulation biases in China from meteorology, chemical composition, and emission sources using an efficient machine learning method

Shuai Wang[1], Mengyuan Zhang[1], Yueqi Gao[1], Peng Wang[2,3], Qingyan Fu[4], Hongliang Zhang[1,3,5]

[1]Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, China

[2]Department of Atmospheric and Oceanic Sciences, and Institute of Atmospheric Sciences, Fudan University, Shanghai, 200438, China

[3]IRDR ICoE on Risk Interconnectivity and Governance on Weather/Climate Extremes Impact and Public Health, Fudan University, Shanghai, China;

[4]Shanghai Environmental Monitoring Center, Shanghai 200235, China

[5]Institute of Eco-Chongming (IEC), Shanghai 200062, China

Correspondence to: Hongliang Zhang (zhanghl@fudan.edu.cn)

Abstract. Chemical transport models (CTMs) are widely used for air pollution modeling, which suffer from significant biases due to uncertainties in simplified parameterization, meteorological fields, and emission inventories. Accurate diagnosis of simulation biases is critical for improvement of models, interpretation of results, and management of air quality, especially for the simulation of fine particulate matter ($PM_{2.5}$). In this study, an efficient method with fast speed and low requirement of computational resources based on the tree-based machine learning (ML) method, the Light Gradient Boosting Machine (LightGBM), was designed to diagnose CTMs simulation biases. The drivers of the Community Multiscale Air Quality (CMAQ) model biases compared to observations in simulating $PM_{2.5}$ concentrations from three perspectives of meteorology, chemical composition, and emission sources. The source-oriented CMAQ was used to diagnose the influences of different emission sources on $PM_{2.5}$ biases. The model can capture the complex relationship between input variables and simulation bias well, meteorology, $PM_{2.5}$ components, and source sectors can partially explain the simulation bias. The CMAQ model underestimates $PM_{2.5}$ by -19.25 to -2.66 $\mu g/m^3$ in 2019, especially in winter and spring and high $PM_{2.5}$ events. Secondary organic components showed the largest contribution to $PM_{2.5}$ simulation bias for different regions and seasons (13.8 - 22.6%) among components. Relative humidity, cloud cover, and soil surface moisture were the main meteorological factors contributing to $PM_{2.5}$ bias in the North China Plain, Pearl River Delta, and Northwestern, respectively. Both primary and secondary inorganic components from residential sources showed the largest contribution to this bias (12.05 % and 12.78 %), implying large uncertainties in this sector. The ML-based methods provide valuable complements to traditional mechanism-based methods for model improvement, with high efficiency and low reliance on prior information.

1 Introduction

Fine particulate matter ($PM_{2.5}$) is a complex mixture of primary particulate matter (PPM) and secondary inorganic/organic components (SIA/SOA), with adverse effects on public health and ecosystems. Ambient levels of $PM_{2.5}$ are influenced by meteorological conditions, emission from different sources, and atmospheric chemical processes (Organization, 2021; Xiao et al., 2021a; Yang et al., 2016; Liu et al., 2021b; Zhai et al., 2019). China has experienced severe $PM_{2.5}$ pollution over the past two decades (Bai et al., 2022; Liang et al., 2020a). For effective air quality management, accurate $PM_{2.5}$ modeling is essential. Chemical transport models (CTMs) like the Community Multiscale Air Quality (CMAQ) model, have been widely developed and applied to $PM_{2.5}$ simulations through atmospheric processes of dispersion, deposition, and chemical reactions (Qiao et al., 2018; Wang et al., 2021; Hu et al., 2017a). Application of CTMs simulations is often limited by the biases due to uncertainties of simplified parameterization, meteorological prediction, emission inventories, and initial and boundary conditions

(Binkowski and Roselle, 2003; Hu et al., 2014; Hu et al., 2016; Wang et al., 2023b; Wang et al., 2021). Thus, it is essential to diagnose specific sources of simulation biases according to specific model applications, including grid resolution, parameterization, mechanisms, meteorological inputs, and emission inventories.

Traditional bias diagnosis approaches for CTM models usually rely on empirical and prior assumptions with extensive sensitivity testing and high demands on computational resources, such as Monte Carlo methods or Latin hypercube sampling (Beekmann and Derognat, 2003; Hanna et al., 2005; Aleksankina et al., 2019). Recently machine learning (ML) methods, like Random Forest and eXtreme Gradient Boosting (XGBoost), have been widely used in environmental science researches due to their simple structure, fast speed, and ability to deal with no-linear relationships (Liu et al., 2022). Many studies used ML to predict air pollutant concentrations like $PM_{2.5}$ and ozone (Wei et al., 2021a; Sun et al., 2021; Zhu et al., 2022; Bai et al., 2022), improve the accuracy of CTMs simulations (Wang et al., 2023b; Wei et al., 2020), and explain the prediction results using interpretable ML techniques (Hou et al., 2022; Li et al., 2023; Stirnberg et al., 2021). To date, few studies have used ML to diagnose the simulation bias of CTMs. A study has shown the potential of machine learning in explaining the simulation bias of ozone (Ye et al., 2022). However, as a complex multi-phase mixture, it is still challenging to diagnose biases in $PM_{2.5}$ simulations using ML methods (Liu and Xing, 2022). Moreover, given the significant impact of emissions, it is instructive to diagnose CTMs biases of $PM_{2.5}$ based on a source-appointment perspective.

In this study, we use LightGBM model, an efficient ensemble ML approach, to diagnose the drivers of CMAQ biases in simulating $PM_{2.5}$ concentrations. A source-oriented version of CMAQ is used to track sectoral source contributions to $PM_{2.5}$. Model biases are diagnosed from multiple perspectives, including meteorology, chemical components, and emission sources.

2 Materials and methods

2.1 Surface $PM_{2.5}$ observations

This study specifically targets the year of 2019 due to the extensive availability of observational data, the reliability of emission inventories, and the absence of COVID-19-related disruptions. Hourly $PM_{2.5}$ observations for 2019 are collected from the China National Environmental Monitoring Centre (CNEMC, http://www.cnemc.cn/). The daily observations data <0.1 % quantile and >99.9 % quantile, $PM_{2.5}$ exceeds PM_{10}, and days with less than 20 valid hourly records are excluded. For observation sites located on the same CMAQ simulation grid (36 km × 36 km), average $PM_{2.5}$ concentrations of these sites were calculated to compare with CMAQ simulation. Approximately 350,000 observations, which met the quality control criteria, were selected from the entire time series data points collected from various monitoring stations. The distribution of observation sites (about 1200) is shown in Figure S1. The stations are unevenly distributed, with dense stations in eastern populated areas and sparse stations in western Xinjiang and Tibet. Analysis has been carried out on several haze-polluted regions and the whole country (Figure S1), including Beijing-Tianjin-Hebei (BTH); the Yangtze River Delta (YRD); the Pearl River Delta (PRD); the Sichuan Basin (SCB); and Northwestern China (NWCHN).

2.2 CMAQ simulation

The CMAQ simulation (36 km×36 km) was carried out to simulate $PM_{2.5}$ components in mainland China and surrounding regions in 2019. The list of $PM_{2.5}$ components simulated by CMAQ is shown in Table A1. The Weather Research & Forecasting Model (WRF v4.2) was used to generate meteorological fields driven by the National Centers for Environmental Prediction (NCEP) FNL Operational Model Global Tropospheric Analyses dataset (http://rda.ucar.edu/datasets/ds083.2/) (Commerce, 2000). Several meteorological factors (Table A1) that are highly relevant to aerosol concentrations are selected for ML model building (Xiao et al., 2021b; Chen et al., 2020b; Meng et al., 2019). The CMAQ v5.0.2 with a modified SAPRC-11 photochemical mechanism and AERO6 aerosol module was applied for aerosol simulations (Carter and Heo, 2013; Ying et al., 2015; Binkowski and Roselle, 2003). The Multi-resolution Emission Inventory for China (MEIC) was used as anthropogenic emission (http://meicmodel.org/), and the Model for Emissions of Gases and Aerosols from Nature (MEGAN) version 2.1 was used to generate biogenic emissions (Guenther et al., 2012; Guenther et al., 2006). The Fire INventory from NCAR (FINN) based on satellite was used to generate open burning emissions (Wiedinmyer et al., 2011).

The source apportionment method was used to quantify the contributions of industry, energy, residential, transportation, agriculture, open burning, and biogenic sources to PPM and SIA through a modified version of CMAQ (Zhang et al., 2012; Ma et al., 2021; Qiao et al., 2018). PPM from different source sectors are tracked by non-reactive tracers (10^{-5} of the PPM emission rates), and source-specific PPM concentrations are then calculated by multiplying the tracer with 10^5. The contributions of source sectors to SIA are quantified using specific reactive tagged tracers. Specifically, NO_x, SO_2, and NH_3 from different sources were tracked separately through a series of chemical and physical processes involved in SIA formation. The source of SOA was not traced due to the complex and currently imperfect mechanism of SOA formation and the high uncertainties in the precursor VOCs emissions (Liu et al., 2021b; Hu et al., 2017b). Details of source apportionment can be found in previous studies (Zhang et al., 2012; Ma et al., 2021; Qiao et al., 2018; Ying et al., 2014). The contributions of source sectors to SOA were not tracked due to insufficient knowledge of its precursors and incomplete formation mechanisms (Yang et al., 2019; Carlton et al., 2007; Zhang et al., 2011).

2.3 Machine learning method

Tree-based ML models typically outperform deep learning approaches in tabular data (e.g., air pollutant observation datasets), and thus have been widely developed (Grinsztajn et al., 2022). Wei et al. (2021a) compared several models when reconstructing $PM_{2.5}$ data records in China and found that the tree model showed superior performance. The LightGBM model is an optimized Gradient Boosting Decision Tree (GBDT) (Ke et al., 2017), and has shown accurate performance in many fields (Wei et al., 2021b; Yan et al., 2021; Sun et al., 2020; Liang et al., 2020b). Compared to XGBoost, a widely used GBDT, LightGBM uses Histogram's decision tree algorithm along with Gradient-based One-Side Sampling (GOSS), which saves memory and computation time (Ke et al., 2017). Three tree-based models, Random Forest, XGBoost, and LightGBM, were compared in our previous study (Wang et al., 2023a). Using the same input data and hyperparameters, LightGBM is as accurate as XGBoost, but faster and less overfitting (the difference in accuracy between training and testing). Besides, Multiple colinearities between features such as pollutant concentrations and meteorological factors can greatly affect the performance of traditional linear models. When independent variables are correlated, changes in one variable are associated with changes in the other, making it difficult for the model to independently estimate the relationship between each independent and dependent variable. However, these collinearities do not affect the performance of tree-based models like Random Forest and LightGBM, because they do not require the assumption of feature independence (Belgiu and Drăguţ, 2016; Chen et al., 2016; Ke et al., 2017). So, the lightGBM model was used to diagnose $PM_{2.5}$ simulation biases in this study. Two metrics were calculated to evaluate model performance, including the coefficient of determination (R^2) and the root mean square error (RMSE) (Wei et al., 2020).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \hat{y})^2} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \tag{2}$$

Cross-validation (5-fold) combined with RMSE was used to select hyperparameters. The dataset was randomly divided into five parts, one was taken in turn as the test set, and the rest was used for training, which was repeated five times, and the average test RMSE was calculated. Looping to increase model complexity, ending the loop and returning to the hyperparameters when the model test RMSE does not decrease significantly (< 0.01) or the gap between training and test RMSE increases significantly (< 0.05). The separate test sets (not involved in the training and CV hyperparameter selection process) were divided by randomly sampling 20% of the data from all stations in the region of interest.

The target variable was defined as the difference between observed and simulated daily $PM_{2.5}$ concentrations, and the key contributors to the simulation bias were identified through the relative importance (calculated by gain) of the input features (Ye et al., 2022; Loyola-González et al., 2023). Three categories of input variables were designed to separately fit the simulation biases: meteorological factors, chemical components, and emission sources. Meteorological factors, including wind speed, wind direction, temperature, humidity, surface pressure, cloud fraction, and boundary layer height, are used to investigate the impact of meteorology on the CMAQ simulation biases. The components of $PM_{2.5}$ are divided into SIA (sulfate, nitrate, ammonium), primary/secondary organic aerosols (POA/SOA), elemental carbon (EC), and other components. The contributions to the simulation bias were quantified using seven sectoral sources: industry, energy, residential, transportation, agriculture, open burning, and biogenic emissions.

3 Results and discussion

3.1 Observation and simulation of $PM_{2.5}$

Figure 1a shows the time series of observed and simulated daily surface $PM_{2.5}$ concentrations in China and five regions (BTH, YRD, PRD, SCB, and NWCHN) over 2019. Observed $PM_{2.5}$ concentrations were highest in BTH (51.172 $\mu g/m^3$) and lowest in PRD (28.273 $\mu g/m^3$). The CMAQ model underestimates $PM_{2.5}$ concentrations of -8.59 $\mu g/m^3$, -2.66 $\mu g/m^3$, -6.21 $\mu g/m^3$, and -19.25 $\mu g/m^3$ in the BTH, YRD, PRD, and NWCHN, respectively (Figure 1b). Moreover, the underestimation occurred mainly in winter and spring (Figure 1c), as well as high $PM_{2.5}$ events (Figure 1d) (Hu et al., 2016; Huang et al., 2017).

Table A2 shows the validation of CMAQ simulations against observations in different regions. Four indicators (MNB: mean normalized bias; MNE: mean normalized error; MFB: mean fractional bias; MFE: mean fractional error) were used to systematically evaluate the performance of the CMAQ simulations. The $PM_{2.5}$ simulations in the BTH, YRD, and PRD regions were in better agreement with observations, with average MNB of -0.08, -0.07, and -0.08 respectively (within the standard of 0.66). The $PM_{2.5}$ simulations in SCB and NWCHN regions show large biases with MNB of 0.46 and -0.42 respectively. The differences of CMAQ performance between regions can be attributed to multiple factors including emission inventories, dominant mechanisms of $PM_{2.5}$ generation, topographic, and meteorology conditions (Ma et al., 2021; Xue et al., 2019; Hu et al., 2014).

Annual and monthly mean $PM_{2.5}$ components (SIA, POA, SOA, EC, and other components) were calculated for China and five key regions (Figure 2). $PM_{2.5}$ and its components show similar spatial distribution, with high concentrations occurring in the eastern regions (SCB, BTH, and central YRD). SOA showed high concentrations in summer over China (6.80 $\mu g/m^3$), which could be related to enhanced solar radiation and atmospheric oxidation capacity in summer (precursors of SOA such as isoprene are highly dependent on temperature and light) (Yang et al., 2019; Chen et al., 2020a; Liu et al., 2021b). Nitrate and POA were the dominant components in winter (10.14 $\mu g/m^3$ and 9.11 $\mu g/m^3$ respectively). In BTH and SCB, POA accounts for a higher proportion than nitrate in winter, whereas nitrate has a higher proportion in YRD. Nitrate showed higher concentration than sulfate in most regions and seasons due to the implementation of coal combustion control policies (Shang et al., 2021; Liu et al., 2021b; Xu et al., 2019).

The results of the $PM_{2.5}$ sectoral source appointment (Figure 3 and Figure S2) show that industries and residential sources were the main contributors to daily $PM_{2.5}$ concentrations for all regions and seasons, with seasonal fractional contributions of 25.31 - 31.92 % and 11.13 - 30.64 %, respectively). The seasonal average fractional contributions from energy, transportation, and agricultural NH_3 in the whole China were 3.26 - 5.67%, 6.82 - 11.26 %, and 7.50 - 8.67 %, respectively. The contributions from biogenic source were negligible in all regions and seasons (< 1 %). In contrast to the contributions from energy, transportation, industrial, and agricultural sources, significant seasonal variations occurred from residential source in all five regions, with high contribution in winter (17.60 - 30.90 %) and low contribution in summer (5.53 - 16.46 %).

As the secondary component makes up a large proportion of the total $PM_{2.5}$, the source sectors of SIA were analyzed for five regions (Figure S2). High concentrations of SIA were found in winter (12.36 - 34.08 $\mu g/m^3$), with large contribution from industrial, agricultural, and transportation sources (31.49 - 36.41 %, 20.40 -22.40 %, and 19.77 - 22.46 %). The low contribution of the residential sector to secondary $PM_{2.5}$ but the high contribution to total $PM_{2.5}$ indicates that most residential emission sources emit PPM directly, with a small fraction of secondary generation. The contributions from biogenic and open burning sectors to SIA were relatively low in all regions and seasons (< 10 %).

3.2 Drivers of $PM_{2.5}$ simulation bias

The ML models were trained separately using meteorology, $PM_{2.5}$ components, and source sectors for different regions and seasons, and separate test sets were used to evaluate the model performance (Figure 4). All three feature combinations can partially explain the simulation bias. The mean test R^2 for meteorology, $PM_{2.5}$ components, and source sectors were 0.64, 0.52, and 0.50, respectively, and the RMSE was 17.41, 19.82, and 19.56 μg/m^3, respectively. The model performed better in summer than in winter. This may be attributed to the high simulation biases in winter due to severe $PM_{2.5}$ pollution and complex causes, while $PM_{2.5}$ pollution in summer is lighter with lower CMAQ simulation bias.

Using $PM_{2.5}$ components as input features to fit the total simulation bias enables the identification of components with large simulation bias. Among the $PM_{2.5}$ components (Figure S4), SOA showed the largest contribution to $PM_{2.5}$ simulation bias for different regions and seasons (13.8 - 22.6%), which is consistent with previous studies (Liu et al., 2021b; Yang et al., 2019; Fry et al., 2014). The inorganic aerosols (e.g. sulfates) are produced mainly by chemical pathways, while SOA is produced by the condensation of numerous partially oxidized gases and is therefore influenced by complex precursor concentrations and multi-stage oxidation processes. The incomplete description of SOA formation pathways in CTMs models (simplified SOA parameterization) leads to significant differences between simulations and observations (Carlton et al., 2007; Zhang et al., 2018; Yang et al., 2019). In addition, biogenic emissions play an important role in SOA formation, with biogenic SOA accounting for more than 70% of total SOA in China during summer (Hu et al., 2017b; Wu et al., 2020), so uncertainties in biogenic emissions can further contribute to uncertainties in SOA. Nitrate showed a large contribution to $PM_{2.5}$ simulation bias in winter at BTH, which is consistent with the previous study (Liu and Xing, 2022). Nitrate contribution to simulation bias further implies the inaccuracy of nitrate simulations, which can relate to the imperfect pathways of nitrate production (e.g., non-homogeneous oxidation) in SAPRC11 (that we used) and the uncertainties of nitrate precursor emission inventories in winter (Xu et al., 2019; Zhang et al., 2018; Carter and Heo, 2013).

The contribution of meteorological factors to the simulation bias varies across regions and seasons (Figure 5). In the BTH region, surface pressure and relative humidity contribute the most to the simulation bias. In the PRD region, relative humidity, cloud cover, and wind direction were the main contributors in all four seasons.

Humidity positively or negatively influences $PM_{2.5}$ concentrations through several mechanisms. By enhancing $PM_{2.5}$ hygroscopic increase, promoting the secondary formation, and facilitating the gas-to-particle partitioning, high humidity positively influences $PM_{2.5}$ concentrations and contributes significantly to haze pollution (Chen et al., 2020b; Cheng et al., 2015; Zhang et al., 2011). The contribution of humidity to CMAQ simulation biases can partly attributed to the uncertainties of WRF simulation. The mean RMSE of relative humidity from WRF simulations versus observations was 20.38% in this study (Table A3). In addition, imperfections in the mechanism of humidity-promoted secondary particle formation (e.g., non-homogeneous catalysis of SOA) can also lead to simulation biases (Zhang et al., 2011; Liu et al., 2021b). Atmospheric pressure indirectly influences $PM_{2.5}$ concentrations through other meteorological factors (e.g., humidity, wind, etc.). High-pressure systems are connected to stationary weather, which is unfavorable for $PM_{2.5}$ dispersion. On the other hand, low pressure is usually accompanied by high humidity, influencing $PM_{2.5}$ nucleation, condensation, and coagulation, leading to increased $PM_{2.5}$ concentrations (Chen et al., 2020b). Therefore, the influence of atmospheric pressure on the CMAQ simulation biases in the BTH region may be attributed to the uncertainties of meteorological field (Bei et al., 2017; Zhang et al., 2015). The contribution of wind direction in YRD may also related to the uncertainties of WRF simulation (mean RMSE: 90.39 °). Aerosols have feedback on meteorology (Qu et al., 2021). In addition to directly changing the radiation received by the earth through scattering and absorbing (direct radiation effect), $PM_{2.5}$ can also influence radiation through aerosol-cloud interactions (indirect radiation effect) (Zhao et al., 2017; Yang et al., 2016). Moreover, $PM_{2.5}$ can act as cloud condensation and nucleation sites, contributing to clouds' microphysical development and precipitation formation process (Wang et al., 2020). However, the aerosol-to-meteorological feedback mechanism is missing in CMAQ used in this study. A previous study showed the dominant role of cloud chemistry in aerosol-cloud interactions in southern China (Zhao et al., 2017). Therefore, the influence of cloud cover on simulation biases in YRD can attributed to the lack of aerosol feedback mechanism.

In the NWCHN region, soil surface moisture and stomatal conductance contributed significantly to the simulation bias. These factors can be associated with ground-level sand rise and dust emission (Liu et al., 2021c). Underestimation of dust aerosol in NWCHN can be attributed to emission (both natural and anthropogenic sources), and an accurate emission inventory (empirical- or physical-based numerical models) should be established in Northwest China by detailed activity data and emission factors (Hu et al., 2016; Liu et al., 2021a). In addition, the parameterization and mechanism for dust aerosol simulation in CMAQ should be further examined and improved.

Dry and wet days were divided to analyze the influence of humidity on the simulation biases (Table A4). In most areas of China, CMAQ underestimates $PM_{2.5}$ more severely on dry days than on wet days, with larger absolute biases (-14.56 $\mu g/m^3$, -7.09 $\mu g/m^3$, -7.11 $\mu g/m^3$, and -27.87 $\mu g/m^3$ in spring, summer, autumn, and winter respectively). In dry days, BTH showed severe underestimation in winter (-22.86 $\mu g/m^3$), while PRD showed large simulation bias in spring (-21.55 $\mu g/m^3$). Severe underestimation of $PM_{2.5}$ was observed in both wet and dry days at NWCHN. These underestimates of $PM_{2.5}$ in dry days can related to the dry deposition process. Dry deposition is a critical but highly uncertain sink for aerosols, which depends on the chemical and physical properties of aerosols, and is influenced by land surface properties and meteorological conditions (Shu et al., 2022). Different land-use types (e.g., vegetation, deserts, and snow) possess markedly different capacities to capture particulate matter. The CMAQ model in this study used the dry deposition scheme PR11 from Pleim and Ran (Pleim and Ran, 2011). This study shows that the PR11 scheme underestimates $PM_{2.5}$ concentrations in China. Recent studies in the United States also showed an underestimation of PM_{10} concentrations (Shu et al., 2022). Therefore, it is necessary to further develop and optimize the dry deposition scheme, especially for $PM_{2.5}$. $PM_{2.5}$ underestimation in wet days may be attributed to the biases of wet deposition and secondary organic aerosol formation under high humidity conditions (Wu et al., 2018; Ryu and Min, 2022; Liu et al., 2021b; Zhang et al., 2011).

Source sector contributions of PPM and SIA (obtained from the source-oriented CMAQ) were used to build the model and diagnose the influences of different emissions sources on $PM_{2.5}$ simulation biases (Figure 6). The PPM and SIA from residential showed the largest contribution (12.05 % and 12.78 %) to $PM_{2.5}$ simulation bias. The same conclusion was obtained when building a model with total $PM_{2.5}$ concentrations from different source sectors (Table A5). $PM_{2.5}$ from residential emissions is the main contributor to the CMAQ simulation bias, accounting for 20.2% of the total bias.

In China, the residential sector consumed fossil fuels (coal, oil, and natural gas) and biofuels (wood and crop straw) with low combustion efficiency for cooking and heating and emitted large amounts of air pollutants (Li et al., 2017). However, due to the lack of reliable data (locally accurate emission factor and fuel consumption data), the residential sector has been recognized as a major uncertainty source in anthropogenic emission inventories (Liu et al., 2021d; Shen et al., 2021), which is consistent with the results identified by machine learning in this study. Therefore, developing an accurate residential sector emissions inventory is essential for accurate $PM_{2.5}$ modeling, which requires reliable data of fuel consumption and emission factors based on fuel type, fuel characteristics, and combustion conditions (Liu et al., 2021d).

3.3 Comparisons and uncertainties

Huang et al. (2019) used a new reduced-form model coupled with a higher-order decoupled direct method and stochastic response surface model to identify sources of uncertainty in CMAQ simulations. An analysis of the PRD of China in the spring of 2013 revealed a systematic underestimation of SOA and identified wind speed and primary $PM_{2.5}$ emissions as key sources of uncertainties in $PM_{2.5}$ simulations, which is consistent with the results identified using LightGBM in this study. Aleksankina et al. (2019) identified $PM_{2.5}$ simulation bias in Europe using optimised Latin hypercube sampling and also demonstrated the important impact of primary emissions on $PM_{2.5}$ simulation uncertainties. Liu and Xing (2022) used a fully connected neural network to identify $PM_{2.5}$ simulation biases and found that NO_2 is the main contributor in BTH during heavily polluted events in the winter, which is consistent with the main contribution of nitrate that we found in the BTH (Figure S4).

Although we filtered the features according to their relative importance and priori knowledge, collinearity still exists among the input features. Multicollinearity among features does not affect the performance of tree-based models like Random Forest and LightGBM (Belgiu and Drăguţ, 2016; Chen et al., 2016; Ke et al., 2017), but the contribution of a single feature might be slightly influenced by other features. Previous studies (Hou et al., 2022; Ye et al., 2022) have used ML to explain the causes of air pollution and model bias, and although there was multicollinearity between the input features they used, they got reliable conclusions, showing the minimal impact of multicollinearity and the reliability of tree-based machine learning methods.

The main objective of this study was to diagnose the contributors to CMAQ simulation biases using machine learning, rather than for prediction. Since meteorology or emissions can only partially explain the simulation bias, a low R^2 is justified when fitting the model with only meteorology or emissions variables (Figure 4). We designed a complementary experiment to measure the impact of the model itself by comparing popular regression models (including multiple linear regression, polynomial regression (degree:2), Random Forest, XGBoost, and LightGBM) with the same features ($PM_{2.5}$ components). All models show similar performance (Table A6), e.g., all models show lower R^2 in the winter in the BTH (0.16 - 0.4), and higher R^2 in the SCB region (0.7 - 0.8). This is side evidence that the low R^2 is more affected by the features than the model itself, as the commonly used regression models can hardly obtain high R^2 with insufficient explanatory features (e.g., chemical component features in winter in BTH). Besides, LightGBM shows comparable accuracy to XGBoost but is faster and shows smaller accuracy gaps between training and testing with less overfitting.

Previous pollution prediction studies based on tree models usually added time-related features to describe the temporal pattern of pollutant changes to further improve the prediction ability, e.g., Wei et al. (2021a) improved the model performance by adding temporal features of day of year and Unix timestamps. However, the inclusion of temporal features cannot provide any useful information about contributors of simulation biases instead it is difficult to attribute them to meteorological or emissions contributions.Therefore, temporal features were not included in our model. Besides, the ML bias diagnosis model constructed in this study is based entirely on local data and some temporal and regional processes influencing $PM_{2.5}$ concentrations are not considered in this study, such as vertical transport, long distance transport, which should be better diagnosed in future work, and the main bias contributors of identified by variable importance are in good agreement with the current findings.

4 Conclusion

Based on artificial intelligence technology, this study systematically diagnoses the possible drivers of biases in $PM_{2.5}$ simulation from three perspectives of meteorology, chemical components, and emission sources. The relative importance of multiple factors helps to understand the sources of simulation bias and the deficiencies of the CMAQ mechanisms. SOA is the main contributor to simulation biases among chemical components. $PM_{2.5}$ is more underestimated in dry weather. Among source sectors, residential contributed the most simulation bias for both PPM and SIA. These results provide valuable information for CMAQ model improvement from SOA and dust aerosol underestimation, meteorological field uncertainties, imperfection of the dry deposition scheme, and inaccurate residential emission inventories. As an efficient bias diagnosis method, machine learning based methods provide valuable complements to traditional mechanism-based methods. This approach also greatly reduces the prior information for diagnosing simulation bias and efficiently identifies the important contributors, so it can be easily extended to other CTMs models as well as other pollutants.

Supporting

Additional descriptions of the study domain, WRF-CMAQ simulation performance, concentrations and biases contribution of $PM_{2.5}$ components and sectoral sources.

Code/Data availability

The data and code are publicly accessible in https://zenodo.org/record/7907626, including machine learning code for diagnosing CMAQ simulation bias and the corresponding training dataset. CMAQ is an open-source chemical transport model developed by the US Environmental Protection Agency, which can be downloaded at https://zenodo.org/record/1079898.

Author contribution

Shuai Wang: Methodology, Software, Writing - original draft. **Mengyuan Zhang**: Software, Validation. **Yueqi Gao**: Data curation, Visualization. **Peng Wang**: Methodology, Writing - reviewing and editing. **Qingyan Fu**: Writing - reviewing and editing. **Hongliang Zhang**: Conceptualization, Supervision, Writing - reviewing and editing.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

Aleksankina, K., Reis, S., Vieno, M., and Heal, M. R.: Advanced methods for uncertainty assessment and global sensitivity analysis of an Eulerian atmospheric chemistry transport model, Atmos. Chem. Phys., 19, 2881-2898, 2019.

Bai, K., Li, K., Ma, M., Li, K., Li, Z., Guo, J., Chang, N. B., Tan, Z., and Han, D.: LGHAP: the Long-term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion, Earth Syst. Sci. Data, 14, 907-927, 10.5194/essd-14-907-2022, 2022.

Beekmann, M. and Derognat, C.: Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric pollution over the Paris area (ESQUIF) campaign, J. Geophys. Res.-Atmos., 108, 2003.

Bei, N., Wu, J., Elser, M., Feng, T., Cao, J., El-Haddad, I., Li, X., Huang, R., Li, Z., Long, X., Xing, L., Zhao, S., Tie, X., Prévôt, A. S. H., and Li, G.: Impacts of meteorological uncertainties on the haze formation in Beijing–Tianjin–Hebei (BTH) during wintertime: a case study, Atmos. Chem. Phys., 17, 14579-14591, 10.5194/acp-17-14579-2017, 2017.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS-J. Photogramm. Remote Sens., 114, 24-31, 2016.

Binkowski, F. S. and Roselle, S. J.: Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description, J. Geophys. Res.-Atmos., 108, 2003.

Carlton, A. G., Turpin, B. J., Altieri, K. E., Seitzinger, S., Reff, A., Lim, H.-J., and Ervens, B.: Atmospheric oxalic acid and SOA production from glyoxal: Results of aqueous photooxidation experiments, Atmos. Environ., 41, 7588-7602, https://doi.org/10.1016/j.atmosenv.2007.05.035, 2007.

Carter, W. P. and Heo, G.: Development of revised SAPRC aromatics mechanisms, Atmos. Environ., 77, 404-414, 2013.

Chen, S., Wang, H., Lu, K., Zeng, L., Hu, M., and Zhang, Y.: The trend of surface ozone in Beijing from 2013 to 2019: Indications of the persisting strong atmospheric oxidation capacity, Atmos. Environ., 242, 117801, 2020a.

Chen, T. Q., Guestrin, C., and Assoc Comp, M.: XGBoost: A Scalable Tree Boosting System, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, Aug 13-17, WOS:000485529800092, 785-794, 10.1145/2939672.2939785, 2016.

Chen, Z. Y., Chen, D. L., Zhao, C. F., Kwan, M. P., Cai, J., Zhuang, Y., Zhao, B., Wang, X. Y., Chen, B., Yang, J., Li, R. Y., He, B., Gao, B. B., Wang, K. C., and Xu, B.: Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism, Environ. Int., 139, 10.1016/j.envint.2020.105558, 2020b.

Cheng, Y., He, K. B., Du, Z. Y., Zheng, M., Duan, F. K., and Ma, Y. L.: Humidity plays an important role in the PM2.5 pollution in Beijing, Environ. Pollut., 197, 68-75, 10.1016/j.envpol.2014.11.028, 2015.

Commerce, N. C. f. E. P. N. W. S. N. U. D. o.: NCEP FNL operational model global tropospheric analyses, continuing from July 1999, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, 2000.

Fry, J. L., Draper, D. C., Barsanti, K. C., Smith, J. N., Ortega, J., Winkler, P. M., Lawler, M. J., Brown, S. S., Edwards, P. M., Cohen, R. C., and Lee, L.: Secondary Organic Aerosol Formation and Organic Nitrate Yield from NO3 Oxidation of Biogenic Hydrocarbons, Environ. Sci. Technol., 48, 11944-11953, 10.1021/es502204x, 2014.

Grinsztajn, L., Oyallon, E., and Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data?, arXiv preprint arXiv:2207.08815, 2022.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmos. Chem. Phys., 6, 3181-3210, 2006.

Guenther, A., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2. 1): an extended and updated framework for modeling biogenic emissions, Geosci. Model Dev., 5, 1471-1492, 2012.

Hanna, S., Russell, A., Wilkinson, J., Vukovich, J., and Hansen, D.: Monte Carlo estimation of uncertainties in BEIS3 emission outputs and their effects on uncertainties in chemical transport model predictions, J. Geophys. Res.-Atmos., 110, 2005.

Hou, L. L., Dai, Q. L., Song, C. B., Liu, B. W., Guo, F. Z., Dai, T. J., Li, L. X., Liu, B. S., Bi, X. H., Zhang, Y. F., and Feng, Y. C.: Revealing Drivers of Haze Pollution by Explainable Machine Learning, Environmental Science & Technology Letters, 9, 112-119, 10.1021/acs.estlett.1c00865, 2022.

Hu, J., Chen, J., Ying, Q., and Zhang, H.: One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system, Atmos. Chem. Phys., 16, 10333-10350, 2016.

Hu, J., Wang, Y., Ying, Q., and Zhang, H.: Spatial and temporal variability of PM2. 5 and PM10 over the North China Plain and the Yangtze River Delta, China, Atmos. Environ., 95, 598-609, 2014.

Hu, J., Huang, L., Chen, M., Liao, H., Zhang, H., Wang, S., Zhang, Q., and Ying, Q.: Premature mortality attributable to particulate matter in China: source contributions and responses to reductions, Environ. Sci. Technol., 51, 9950-9959, 2017a.

Hu, J., Wang, P., Ying, Q., Zhang, H., Chen, J., Ge, X., Li, X., Jiang, J., Wang, S., and Zhang, J.: Modeling biogenic and anthropogenic secondary organic aerosol in China, Atmos. Chem. Phys., 17, 77-92, 2017b.

Huang, Z., Zheng, J., Ou, J., Zhong, Z., Wu, Y., and Shao, M.: A Feasible Methodological Framework for Uncertainty Analysis and Diagnosis of Atmospheric Chemical Transport Models, Environ. Sci. Technol., 53, 3110-3118, 10.1021/acs.est.8b06326, 2019.

Huang, Z., Hu, Y., Zheng, J., Yuan, Z., Russell, A. G., Ou, J., and Zhong, Z.: A New Combined Stepwise-Based High-Order Decoupled Direct and Reduced-Form Method To Improve Uncertainty Analysis in PM(2.5) Simulations, Environ Sci Technol, 51, 3852-3859, 10.1021/acs.est.6b05479, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 2017.

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., and He, K.: Anthropogenic emission inventories in China: a review, National Science Review, 4, 834-866, 10.1093/nsr/nwx150, 2017.

Li, T., Zhang, Q., Peng, Y., Guan, X., Li, L., Mu, J., Wang, X., Yin, X., and Wang, Q.: Contributions of Various Driving Factors to Air Pollution Events: Interpretability Analysis from Machine Learning Perspective, Environ. Int., 107861, 2023.

Liang, F., Xiao, Q., Huang, K., Yang, X., Liu, F., Li, J., Lu, X., Liu, Y., and Gu, D.: The 17-y spatiotemporal trend of PM2. 5 and its mortality burden in China, Proc. Natl. Acad. Sci. U. S. A., 117, 25601-25608, 2020a.

Liang, W., Luo, S., Zhao, G., and Wu, H.: Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms, Mathematics, 8, 765, 2020b.

Liu, J. and Xing, J.: Identifying Contributors to PM2.5 Simulation Biases of Chemical Transport Model Using Fully Connected Neural Networks, Journal of Advances in Modeling Earth Systems, 15, 2022.

Liu, J., Ding, J., Rexiding, M., Li, X., Zhang, J., Ran, S., Bao, Q., and Ge, X.: Characteristics of dust aerosols and identification of dust sources in Xinjiang, China, Atmos. Environ., 262, 118651, https://doi.org/10.1016/j.atmosenv.2021.118651, 2021a.

Liu, J., Chu, B., Chen, T., Zhong, C., Liu, C., Ma, Q., Ma, J., Zhang, P., and He, H.: Secondary organic aerosol formation potential from ambient air in Beijing: effects of atmospheric oxidation capacity at different pollution levels, Environ. Sci. Technol., 55, 4565-4572, 2021b.

Liu, S., Xing, J., Sahu, S. K., Liu, X., Liu, S., Jiang, Y., Zhang, H., Li, S., Ding, D., Chang, X., and Wang, S.: Wind-blown dust and its impacts on particulate matter pollution in Northern China: current and future scenarios, Environ. Res. Lett., 16, 114041, 10.1088/1748-9326/ac31ec, 2021c.

Liu, X., Lu, D., Zhang, A., Liu, Q., and Jiang, G.: Data-driven machine learning in environmental pollution: Gains and problems, Environ. Sci. Technol., 56, 2124-2133, 2022.

Liu, X., Shen, G., Chen, L., Qian, Z., Zhang, N., Chen, Y., Chen, Y., Cao, J., Cheng, H., Du, W., Li, B., Li, G., Li, Y., Liang, X., Liu, M., Lu, H., Luo, Z., Ren, Y., Zhang, Y., Zhu, D., and Tao, S.: Spatially Resolved Emission Factors to Reduce Uncertainties in Air Pollutant Emission Estimates from the Residential Sector, Environ. Sci. Technol., 55, 4483-4493, 10.1021/acs.est.0c08568, 2021d.

Loyola-González, O., Ramírez-Sáyago, E., and Medina-Pérez, M. A.: Towards improving decision tree induction by combining split evaluation measures, Knowledge-Based Systems, 277, 110832, https://doi.org/10.1016/j.knosys.2023.110832, 2023.

Ma, J., Shen, J., Wang, P., Zhu, S., Wang, Y., Wang, P., Wang, G., Chen, J., and Zhang, H.: Modeled changes in source contributions of particulate matter during the COVID-19 pandemic in the Yangtze River Delta, China, Atmos. Chem. Phys., 21, 7343-7355, 2021.

Meng, C., Cheng, T. H., Gu, X. F., Shi, S. Y., Wang, W. N., Wu, Y., and Bao, F. W.: Contribution of meteorological factors to particulate pollution during winters in Beijing, Science of the Total Environment, 656, 977-985, 10.1016/j.scitotenv.2018.11.365, 2019.

Organization, W. H.: WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary, 2021.

Pleim, J. and Ran, L.: Surface flux modeling for air quality applications, Atmosphere, 2, 271-302, 2011.

Qiao, X., Ying, Q., Li, X., Zhang, H., Hu, J., Tang, Y., and Chen, X.: Source apportionment of PM2.5 for 25 Chinese provincial capitals and municipalities using a source-oriented Community Multiscale Air Quality model, Science of the Total Environment, 612, 462-471, 10.1016/j.scitotenv.2017.08.272, 2018.

Qu, Y., Voulgarakis, A., Wang, T., Kasoar, M., Wells, C., Yuan, C., Varma, S., and Mansfield, L.: A study of the effect of aerosols on surface ozone through meteorology feedbacks over China, Atmos. Chem. Phys., 21, 5705-5718, 10.5194/acp-21-5705-2021, 2021.

Ryu, Y.-H. and Min, S.-K.: Improving Wet and Dry Deposition of Aerosols in WRF-Chem: Updates to Below-Cloud Scavenging and Coarse-Particle Dry Deposition, Journal of Advances in Modeling Earth Systems, 14, e2021MS002792, https://doi.org/10.1029/2021MS002792, 2022.

Shang, D., Peng, J., Guo, S., Wu, Z., and Hu, M.: Secondary aerosol formation in winter haze over the Beijing-Tianjin-Hebei Region, China, Frontiers of Environmental Science & Engineering, 15, 1-13, 2021.

Shen, H., Luo, Z., Xiong, R., Liu, X., Zhang, L., Li, Y., Du, W., Chen, Y., Cheng, H., Shen, G., and Tao, S.: A critical review of pollutant emission factors from fuel combustion in home stoves, Environ. Int., 157, 106841, https://doi.org/10.1016/j.envint.2021.106841, 2021.

Shu, Q., Murphy, B., Schwede, D., Henderson, B. H., Pye, H. O. T., Appel, K. W., Khan, T. R., and Perlinger, J. A.: Improving the particle dry deposition scheme in the CMAQ photochemical modeling system, Atmos. Environ., 289, 119343, https://doi.org/10.1016/j.atmosenv.2022.119343, 2022.

Stirnberg, R., Cermak, J., Kotthaus, S., Haeffelin, M., Andersen, H., Fuchs, J., Kim, M., Petit, J. E., and Favez, O.: Meteorology-driven variability of air pollution (PM1) revealed with explainable machine learning, Atmos. Chem. Phys., 21, 3919-3948, 10.5194/acp-21-3919-2021, 2021.

Sun, H., Shin, Y. M., Xia, M., Ke, S., Wan, M., Yuan, L., Guo, Y., and Archibald, A. T.: Spatial Resolved Surface Ozone with Urban and Rural Differentiation during 1990–2019: A Space–Time Bayesian Neural Network Downscaler, Environ. Sci. Technol., 56, 7337-7349, 2021.

Sun, X., Liu, M., and Sima, Z.: A novel cryptocurrency price trend forecasting model based on LightGBM, Finance Research Letters, 32, 101084, 2020.

Wang, P., Qiao, X., and Zhang, H.: Modeling PM2.5 and O3 with aerosol feedbacks using WRF/Chem over the Sichuan Basin, southwestern China, Chemosphere, 254, 126735, https://doi.org/10.1016/j.chemosphere.2020.126735, 2020.

Wang, S., Wang, P., Zhang, R., Meng, X., Kan, H., and Zhang, H.: Estimating particulate matter concentrations and meteorological contributions in China during 2000–2020, Chemosphere, 330, 138742, https://doi.org/10.1016/j.chemosphere.2023.138742, 2023a.

Wang, S., Wang, P., Qi, Q., Wang, S., Meng, X., Kan, H., Zhu, S., and Zhang, H.: Improved estimation of particulate matter in China based on multisource data fusion, Science of The Total Environment, 161552, 2023b.

Wang, S. Y., Zhang, Y. L., Ma, J. L., Zhu, S. Q., Shen, J. Y., Wang, P., and Zhang, H. L.: Responses of decline in air pollution and recovery associated with COVID-19 lockdown in the Pearl River Delta, Science of the Total Environment, 756, 10.1016/j.scitotenv.2020.143868, 2021.

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

Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., Li, R., and Cribb, M.: Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM), Atmos. Chem. Phys., 21, 7863-7880, 10.5194/acp-21-7863-2021, 2021b.

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

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J., and Soja, A. J.: The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning, Geosci. Model Dev., 4, 625-641, 10.5194/gmd-4-625-2011, 2011.

Wu, K., Yang, X. Y., Chen, D., Gu, S., Lu, Y. Q., Jiang, Q., Wang, K., Ou, Y. H., Qian, Y., Shao, P., and Lu, S. H.: Estimation of biogenic VOC emissions and their corresponding impact on ozone and secondary organic aerosol formation in China, Atmospheric Research, 231, 10.1016/j.atmosres.2019.104656, 2020.

Wu, Y., Liu, J., Zhai, J., Cong, L., Wang, Y., Ma, W., Zhang, Z., and Li, C.: Comparison of dry and wet deposition of particulate matter in near-surface waters during summer, PloS one, 13, e0199241, 2018.

Xiao, Q., Geng, G., Xue, T., Liu, S., Cai, C., He, K., and Zhang, Q.: Tracking PM 2.5 and O 3 Pollution and the Related Health Burden in China 2013–2020, 2021a.

Xiao, Q., Zheng, Y., Geng, G., Chen, C., Huang, X., Che, H., Zhang, X., He, K., and Zhang, Q.: Separating emission and meteorological contributions to long-term PM2.5 trends over eastern China during 2000–2018, Atmos. Chem. Phys., 21, 9475-9496, 10.5194/acp-21-9475-2021, 2021b.

Xu, Q., Wang, S., Jiang, J., Bhattarai, N., Li, X., Chang, X., Qiu, X., Zheng, M., Hua, Y., and Hao, J.: Nitrate dominates the chemical composition of PM2. 5 during haze event in Beijing, China, Science of the Total Environment, 689, 1293-1303, 2019.

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

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X.: LightGBM: accelerated genomically designed crop breeding through ensemble learning, Genome Biology, 22, 1-24, 2021.

Yang, W., Li, J., Wang, W., Li, J., Ge, M., Sun, Y., Chen, X., Ge, B., Tong, S., Wang, Q., and Wang, Z.: Investigating secondary organic aerosol formation pathways in China during 2014, Atmos. Environ., 213, 133-147, https://doi.org/10.1016/j.atmosenv.2019.05.057, 2019.

Yang, X., Zhao, C. F., Guo, J. P., and Wang, Y.: Intensification of aerosol pollution associated with its feedback with surface solar radiation and winds in Beijing, J. Geophys. Res.-Atmos., 121, 4093-4099, 10.1002/2015jd024645, 2016.

Ye, X., Wang, X., and Zhang, L.: Diagnosing the Model Bias in Simulating Daily Surface Ozone Variability Using a Machine Learning Method: The Effects of Dry Deposition and Cloud Optical Depth, Environ. Sci. Technol., 56, 16665-16675, 10.1021/acs.est.2c05712, 2022.

Ying, Q., Li, J., and Kota, S. H.: Significant contributions of isoprene to summertime secondary organic aerosol in eastern United States, Environ. Sci. Technol., 49, 7834-7842, 2015.

Ying, Q., Wu, L., and Zhang, H.: Local and inter-regional contributions to PM2. 5 nitrate and sulfate in China, Atmos. Environ., 94, 582-592, 2014.

Zhai, S., Jacob, D. J., Wang, X., Shen, L., Li, K., Zhang, Y., Gui, K., Zhao, T., and Liao, H.: 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.

Zhang, H., Surratt, J. D., Lin, Y. H., Bapat, J., and Kamens, R. M.: Effect of relative humidity on SOA formation from isoprene/NO photooxidation: enhancement of 2-methylglyceric acid and its corresponding oligoesters under dry conditions, Atmos. Chem. Phys., 11, 6411-6424, 10.5194/acp-11-6411-2011, 2011.

Zhang, H., Li, J., Ying, Q., Yu, J. Z., Wu, D., Cheng, Y., He, K., and Jiang, J.: Source apportionment of PM2. 5 nitrate and sulfate in China using a source-oriented chemical transport model, Atmos. Environ., 62, 228-242, 2012.

Zhang, Q., Quan, J., Tie, X., Li, X., Liu, Q., Gao, Y., and Zhao, D.: Effects of meteorology and secondary particle formation on visibility during heavy haze events in Beijing, China, Science of the Total Environment, 502, 578-584, 2015.

Zhang, R., Sun, X. S., Shi, A. J., Huang, Y. H., Yan, J., Nie, T., Yan, X., and Li, X.: Secondary inorganic aerosols formation during haze episodes at an urban site in Beijing, China, Atmos. Environ., 177, 275-282, 10.1016/j.atmosenv.2017.12.031, 2018.

Zhao, B., Liou, K.-N., Gu, Y., Li, Q., Jiang, J. H., Su, H., He, C., Tseng, H.-L. R., Wang, S., Liu, R., Qi, L., Lee, W.-L., and Hao, J.: Enhanced PM2.5 pollution in China due to aerosol-cloud interactions, Sci Rep, 7, 4453, 10.1038/s41598-017-04096-8, 2017.

Zhu, Q., Bi, J., Liu, X., Li, S., Wang, W., Zhao, Y., and Liu, Y.: Satellite-based long-term spatiotemporal patterns of surface ozone concentrations in China: 2005–2019, Environ. Health Perspect., 130, 027004, 2022.

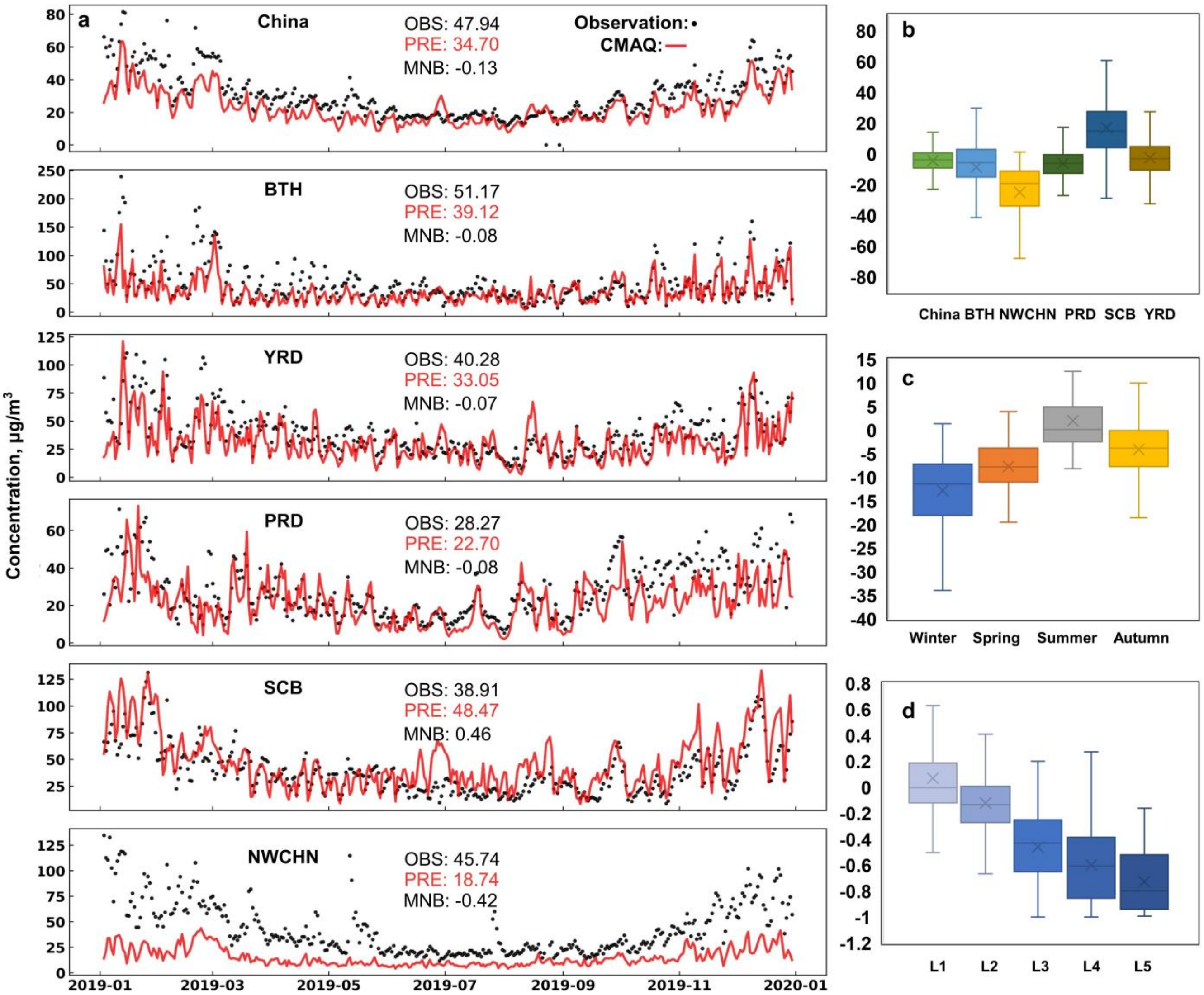


Figure 1. (a) The time series of observed (black) and CMAQ simulated (red) daily surface PM$_{2.5}$ concentrations in China and five regions. Mean concentrations of the observed and simulated PM$_{2.5}$ and MNB also shown in the inset. (b) Box plots of CMAQ simulated biases (simulated minus observed) for different regions. Crosses indicate average values and outliers are determined to be > 1.5 times of the upper quartile and < 1.5 times of the lower quartile. (c) Same as (b) but for four seasons. Spring, summer, autumn and winter are defined as March to May, June to August, September to November, December January and February, respectively. (d) Same as (b) but for different PM$_{2.5}$ concentration levels (L1: [0, 35], L2: [35, 75], L3: [75, 115], L4: [115, 150], L5: [150, 1000]).

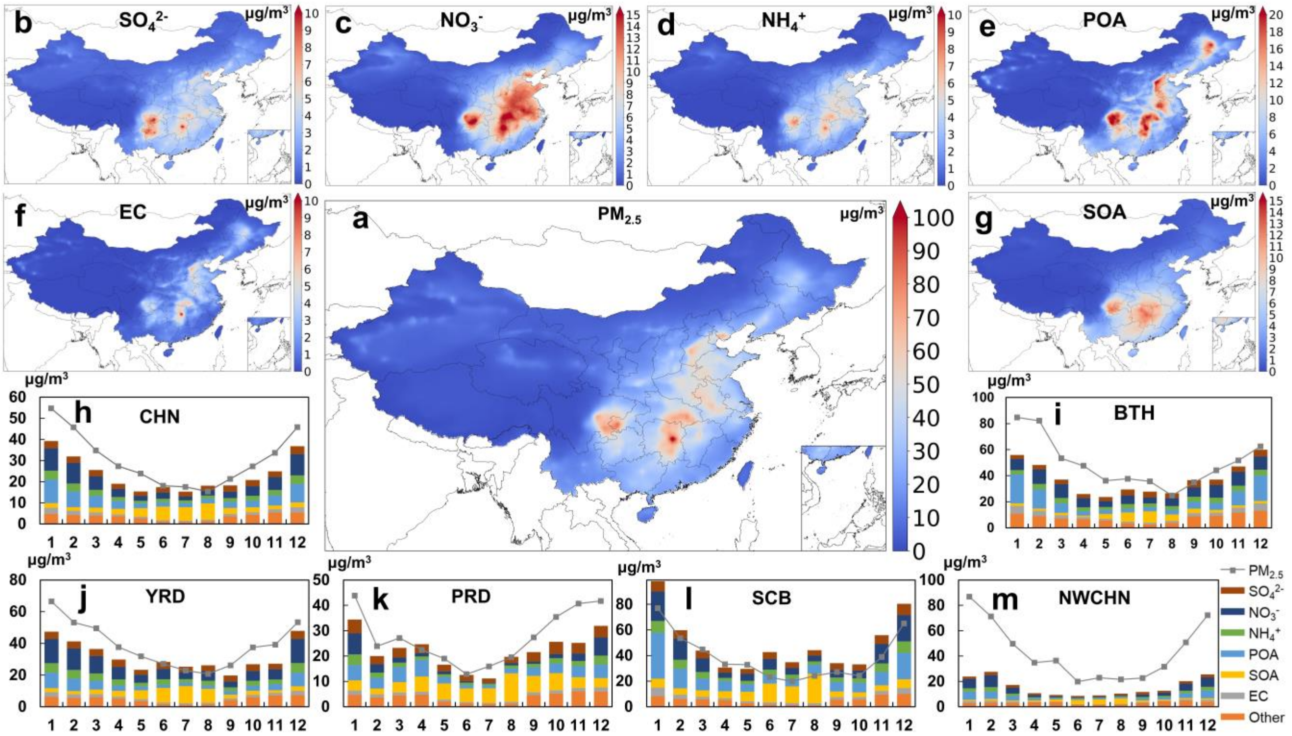


Figure 2. Annual mean concentrations map (a - g) and monthly mean concentrations (h-m) of PM2.5 and its components (SIA , POA, SOA, EC, and other components) for China and five key regions in 2019. Dotted lines in h-m indicate PM2.5 observations

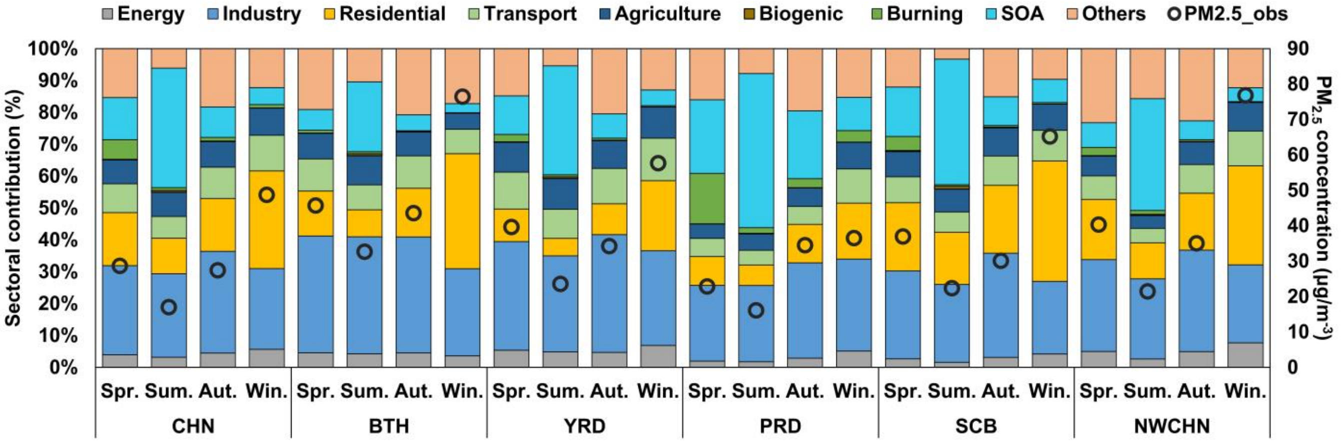
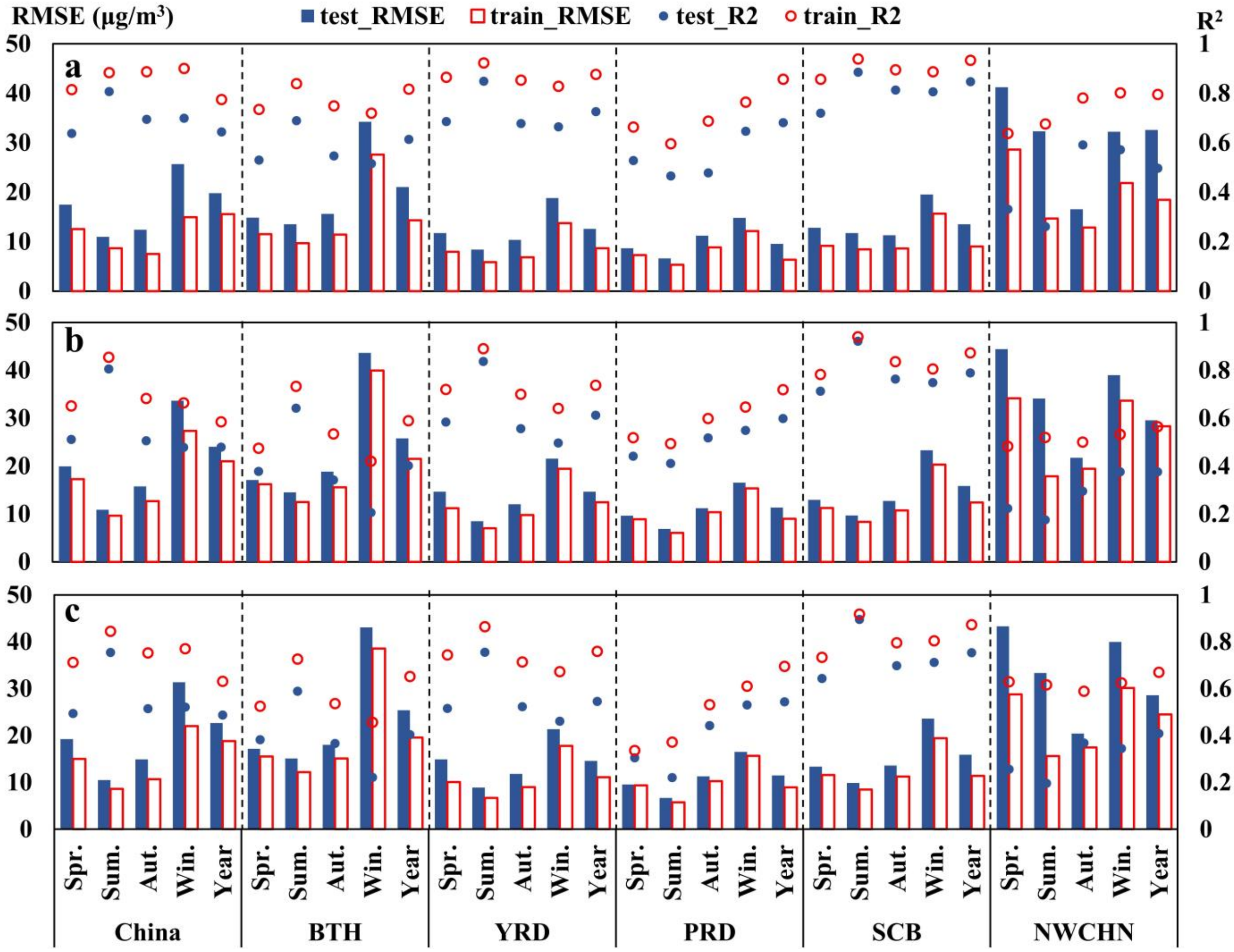


Figure 3. Seasonal average fractional contributions from different sources to PM2.5 concentrations (black circle on the right-hand axis) in China and five regions.



515

Figure 4. Test results of CMAQ bias model training by meteorology (a), $PM_{2.5}$ components (b), and source sectors (c). RMSE unit: μg/m³.

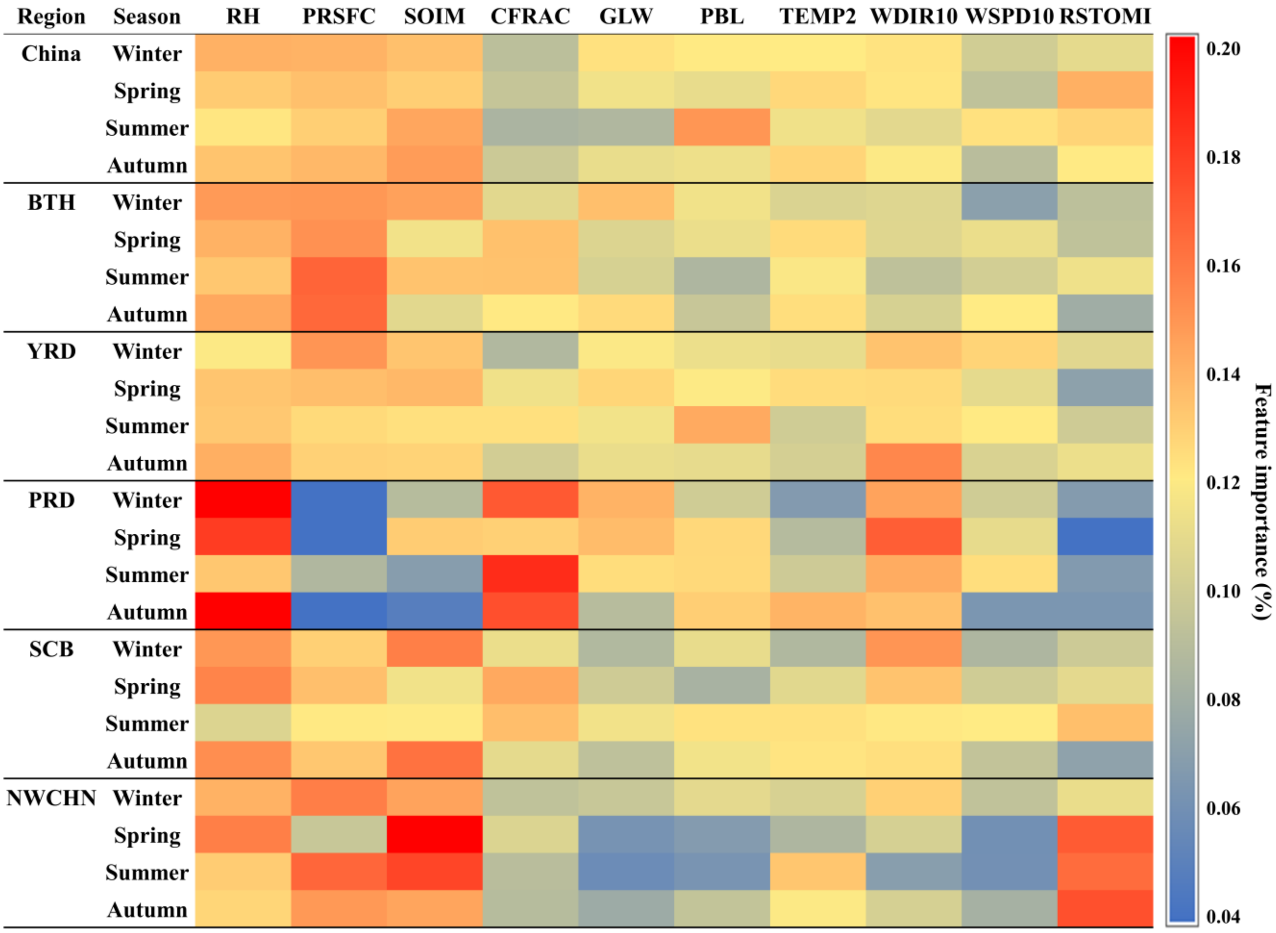


Figure 5. Contribution (%) of each meteorological factor to CMAQ simulation biases by region and season.

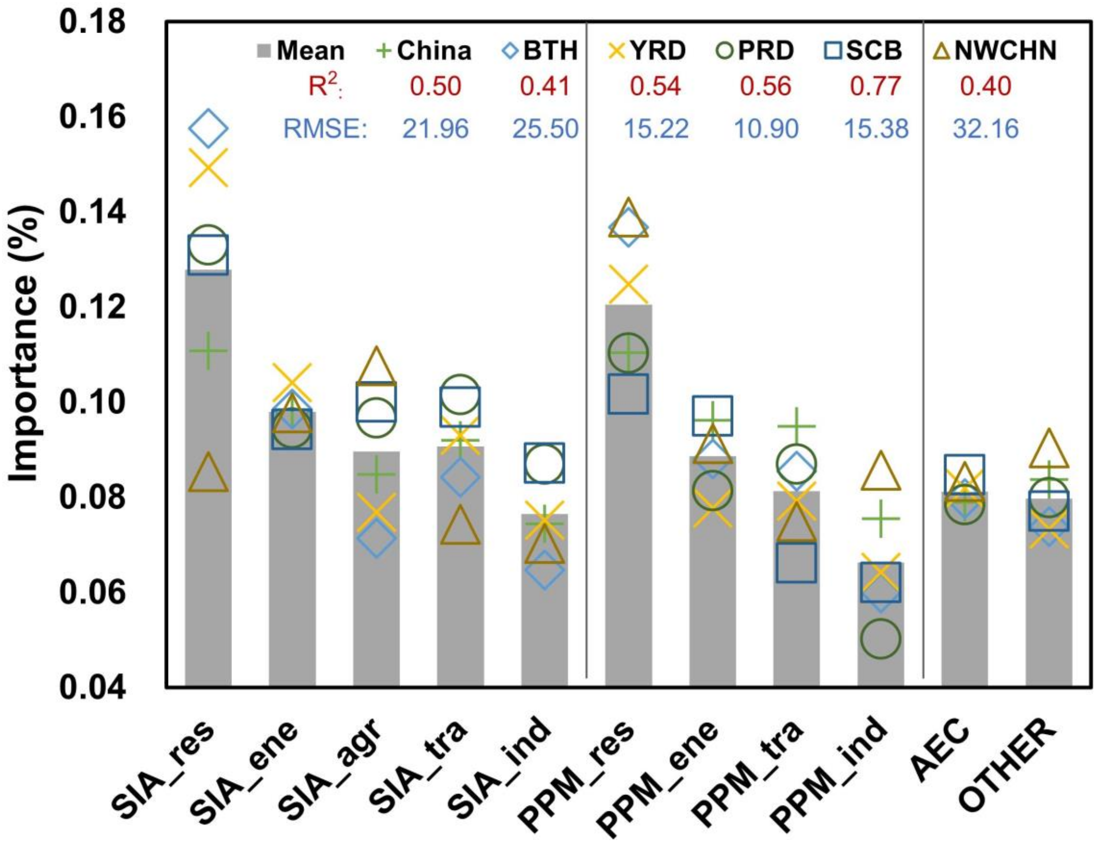


Figure 6. Contribution (%) of each source sectors to CMAQ biases by region and season. res: residential, ene: energy, tra: transportation, agr: agriculture, ind: industry, AEC: elemental carbon, Other: other components.