Explainable ensemble machine learning revealing enhanced anthropogenic emissions of particulate nitro-aromatic compounds in eastern China

Min Li[1], Xinfeng Wang[1,*], Tianshuai Li[1,2], Yujia Wang[1], Yueru Jiang[1,2], Yujiao Zhu[1], Wei Nie[2], Rui Li[3], Jian Gao[4], Likun Xue[1], Qingzhu Zhang[1], Wenxing Wang[1]

[1]Academician Workstation for Big Data in Ecology and Environment, Environment Research Institute, Shandong University, Qingdao, Shandong 266237, China
[2]Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing 210023, China
[3]School of Public Health, MOE Key Laboratory of Coal Environmental Pathogenicity and Prevention, Shanxi Medical University, Taiyuan 030001, China
[4]State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Science, Beijing 100012, China

Correspondence to: Xinfeng Wang (xinfengwang@sdu.edu.cn)

Contents of this files:

Supporting Text Section: S1–S4

Supporting Table S1 and S2

Supporting Figure S1–S8

Number of pages: 20

Text S1. Site description and online instruments

Field observations were performed at eleven sites in eastern China, including four urban sites in Jinan, Guangzhou, Nanjing, and Beijing, five rural sites in Dongying, Wangdu, Yucheng, and Qingdao (including two sampling sites: Qingdao Campus of Shandong University and Entrepreneurship Center of Blue Silicon Valley), and two mountain sites at Mount Tai and Mount Lao (seen in Figure 1). Detailed sampling site information and online measurements are available below.

The Jinan site is situated at the Urban Atmospheric Environment Observation Station (~22 m above ground level) of Shandong University in Jinan, Shandong Province. Jinan, a major industrialized city in North China, has a sampling site characterized by intensive traffic, commercial and residential activities nearby, and extensive industrial facilities. Trace gases, including SO_2, NO_2, and O_3, were monitored with online gas analyzers (Thermo Electronic Corporation, TEC, Model 43C, 42C, and 49C, respectively) and meteorological data were recorded by an automatic meteorological station (CAWS600, Huayun, China). Details about this site were given by Wang et al. (2017b).

The southernmost Guangzhou site is located at the Guangzhou Institute of Geochemistry, Chinese Academy of Sciences in Guangzhou, Guangdong Province. This site is surrounded by education and residential districts, with two heavily trafficked expressways nearby. Related site information was provided by Bi et al. (2016).

The Nanjing site is situated at the Station for Observing Regional Processes of the Earth System (SORPES) in the Xianlin Campus of Nanjing University in Nanjing, Jiangsu Province. Nanjing is a megacity city that dominated by tertiary industries such as finance and software. This site is less influenced by industrial emissions in the vicinity but it is adjacent to the G25 Freeway (~300 m) and G312 National Road (~1.8 km), which may potentially affect the air pollution levels at the sampling location. A more detailed description of this station can be found in a previous study by Ding et al. (2013).

The Beijing site is located at the Chinese Research Academy of Environmental Sciences (CRAES), an urban site with education and residential districts and heavy traffic. As described by Ren et al. (2021), this area was significantly affected by anthropogenic activities and direct emissions. Meteorological parameters, as well as gaseous tracers, were determined simultaneously by employing automated instruments (Chinese Research Academy of Environmental Sciences Supersite for Urban Air Comprehensive Observation and Research).

The Dongying site, where $PM_{2.5}$ samples were collected, is situated at the Yellow River Delta Ecology Research Station of Coastal Wetland, Chinese Academy of Sciences. It is a typical rural site that located at the mouth of the Yellow River, characterized by minimal local anthropogenic emissions. Related tracer gases of SO_2, NO_2, and O_3 were measured by Model 43C (TEC), Model T500U (Teledyne Advanced Pollution Instrumentation, API), and Model 49C (TEC) analyzers, respectively. Meteorological data were also measured online (JZYG, PC-4, China). Detailed information on this site was given by Zhang et al. (2019).

The Wangdu site is located in a rural area of Baoding, Hebei province. The immediate vicinity (within 5 km) of the sampling site consists predominantly of agricultural land. However, this site is affected by anthropogenic emissions from nearby urban cities, such as Beijing, Tianjin, and Shijiazhuang. Trace gases of NO_2 and O_3 were monitored online using a Model 42i analyzer and a Model 49i analyzer (TEC), respectively, while SO_2 was determined by a pulsed UV fluorescence analyzer. Moreover, meteorological parameters were measured using a weather station. More information on the site can be found in Tham et al. (2016).

The measurements conducted at the Yucheng site, situated at the Chinese Academy of Sciences Comprehensive Station, Dezhou, Shandong province. The sampling site is surrounded by agricultural land, but there is the G308 highway located 1.5 km south of the site. Trace gases, including NO_2, O_3, and SO_2 were detected online with Model 42C, 49C, 43C analyzer (TEC), respectively. Data on meteorological parameters were provided by an automatic meteorological station (Model MILOS520, Vaisala, Finland). And details about the site were described by Yao et al. (2016).

The two sampling sites situated in coastal areas in Qingdao are Qingdao Campus of Shandong University and Entrepreneurship Center of Blue Silicon Valley. The two sites are in close proximity to each other, with a linear distance of only 2.2 km (shown in Fig. 1b). They are surrounded by educational and residential districts, villages, and farmlands. As typical rural coastal areas, the two sites are influenced by both anthropogenic and natural sources. Specially, the concentrations of SO_2, NO_2, and O_3 were measured in real time by gas analyzers (Model 43i, 42i, and 49i, respectively). More information on this sampling site can be seen in our previous study (Liu et al., 2022).

The measurement site located on Mount Tai in Tai'an city, Shandong Province, is the highest point in the Northern China Plain, making it an ideal place for studying the transport, sources, and formation processes of air pollutants in northern China. This mountaintop lacks significant local anthropogenic emissions but is influenced by air masses transport processes in the region. Trace gases were recorded using online gas analyzers (Model 43C, Model T200/T500U, and Model T400U for SO_2, NO_2, and O_3, respectively), and meteorological data were obtained from Taishan National Reference Climatological Station. Detailed descriptions of this site were given by Wang et al. (2017c).

The sampling site on Mount Lao is situated in the southeastern part of the Shandong Peninsula in Shandong Province, with a straight-line distance of about 1 km from the coastline. This site is adjacent to a dedicated tourist road and surrounded by several villages, restaurants, and guesthouses. Therefore, Mount Lao is an ideal location for studying the impact of land-sea exchanges on atmospheric pollution characteristics in coastal regions at different scales. In addition, online gas analyzers (Thermo Scientific, U.S.A,) were used to determine the concentrations of trace gases, and meteorological parameters were measured by an ultrasonic automatic weather station (RS-FSXCS-N01-1, China).

Text S2. Analytical method of NACs

$PM_{2.5}$ filter samples were extracted either ultrasonically or using a thermostatic orbital shaker, with methanol containing 30 μL saturated EDTA solution three times for 30 min under a constant temperature condition of 18°C and settled for more than 12 h. Then, the extracts were filtered through a 0.20 μm PTFE membrane syringe filter to remove insoluble impurity. The resulting clear filtrate was evaporated and concentrated with a gentle stream of nitrogen. Finally, the residue was reconstituted to a final volume of 300 μL with methanol containing the internal standard (100 ng mL^{-1} 4-nitrophenol-2,3,5,6-d_4 used for mountain sites and rural Qingdao, 200 ng mL^{-1} 2,4,6-trinitrophenol used for the remaining sites) for further qualitative and quantitative analysis.

NACs in the extracts were then analyzed by using UHPLC-MS equipped with ESI source. The separation of different NACs (only for mountain sites and rural Qingdao) was performed with an Acquity UPLC HSS T3 column (2.1 mm × 100 mm, 1.8 μm particle size, 100Å, Waters, U.S.A.) with a VanGuard column (HSS T3, 1.8 μm) at a flow rate of 0.19 mL min^{-1}. The mobile phase contained eluent A (ultrapure water with 0.1% acetic acid) and eluent B (methanol with 0.1% acetic acid). The gradient program was set as follows: eluent A was initially 99% and kept at 99% for 2.7 min, then gradually decreased to 46% with 12.5 min and kept at 46% for 1 min, and then decreased to 10% with 7.5 min and held for 0.2 min. After that, eluent A increased to 99% in 1.8 min and kept at 99% for the last 17.3 min before the next sample solution. For the remaining sampling sites, the NACs were separated using an Atlantis T3 C18 column (2.1 mm × 150 mm, 3.0 μm particle size, 100Å, Waters, U.S.A.) coupled with a VanGuard column (Atlantis T3, 3.0 μm) at a flow rate of 0.2 mL min-1. The mobile phase consisted of 11% acetonitrile and 0.1% formic acid in ultrapure water (eluent A) and 11% acetonitrile in methanol (eluent B). The proportion of eluent A started with 66%, and then decreased to 44% within 19 min and was kept at 44% for 4min. Finally, it returned to 66% for the last 8 min. The blank samples were extracted and analyzed in the same procedure.

The ESI source was operated in negative mode and eight mass-to-charge ratios including 138, 152, 154, 166, 168, 182, 183, and 197 amu were monitored in real time. Then target NACs were then identified by comparing individual retention times and mass spectra with standard mixtures: NPs (4-nitrophenol (4NP), 3-methyl-4nitrophenol (3M4NP), 2-methyl-4-nitrophenol (2M4NP), and 2.6-dimethyl-4-nitrophenol (2,6DM4NP)), NCs (4-nitrocatechol (4NC), 4-methyl-5-nitrocatechol (4M5NC), 3-methyl-6-nitrocatechol (3M6NC), and 3-methyl-5-nitrocatechol (3M5NC)), NSAs (5-nitrosalicylic acid (5NSA) and 3-nitrosalicylic acid (3NSA)), and DNPs (2,4-dinitrophenol (2,4DNP) and 4-methyl-2,6-dinitrophenol (4M2,6DNP)). Finally, the twelve NACs were quantified using multi-point standard curves ($R^2 > 0.99$) based on gradient standard mixtures. Furthermore, in this study, all reported data in the sample filters were blank-corrected.

Text S3. Positive Matrix Factorization (PMF) analysis

To obtain the potential factor profiles and contributions on NACs, in this study, two to six factors were tested for calculation and evaluation. The difference between Q_{true} provided by the model and calculated Q_{robust}, which calculated by the following Eq. (S1), is used to determine the optimal number of factors for the calculation (Hong et al., 2022; Wu et al., 2020):

$$Q_{\mathrm{robust}} = m \times n - p\,(m \times n) \qquad \text{(S1)}$$

where *m* is the input sample numbers, *n* refers to the number of input species, and *p* refers to the number of factors. The changes in the $Q_{\mathrm{true}}/Q_{\mathrm{robust}}$ ratio values for PMF solutions with 2~6 factors are shown in the Figure S1a. The $Q_{\mathrm{true}}/Q_{\mathrm{robust}}$ value decreased slowly after four factors, so a four-factor solution was chosen as best choice. Besides, the R^2 value between observed NACs and predicted NACs by PMF model was obtained as 0.96 (Fig. S1b), indicating that the PMF analysis is a well-fitting model in this study.

As shown in Figure S7, the major contributions of factor 1 were 4NP (84.9%), 3M4NP (73.7%), and 2M4NP (85.2%). As reported by Lu et al. (2019a), remarkable amounts of NPs were detected in particles from residential coal combustion plumes, with emission factors ranging from 0.01 to 0.94 mg kg^{-1}. Therefore, this factor was defined as coal combustion (CC) activities and the its contribution to NACs was 41.0%.

Factor 2 is featured with the highest loading and contribution (87.3%) of NO_2 and is determined as traffic emissions (TE). Previous studies have indicated that NACs can be directly emitted from traffic activities, with emission factors to be 0.68-89.61 μg km^{-1} (Tremp et al., 1993; Schauer et al., 2002; Lu et al., 2019b), due to the hydrocarbons, polycyclic aromatic hydrocarbons and nitro-polycyclic aromatic hydrocarbons fuel combusting in the engine (Zhang et al., 2014; Cao et al., 2017). The contribution of traffic emission to the concentration of NACs was 29.7%.

Factor 3 is characterized by high contributions of 4NC (56.6%), 4M5NC (84.7%), and 3M6NC (83.9%), which are significant tracers for biomass burning smoke (Iinuma et al., 2010; Claeys et al., 2012), and thus this factor is confirmed as biomass burning (BB). This factor has been considered to be an important source of NACs in recent years that mainly produced by the pyrolysis of lignin (Simoneit et al., 2007). The emission factors of fine NACs from biomass burning were estimated to be 0.75-11.1 mg kg^{-1} (Wang et al., 2017a). The contribution of biomass burning to the NACs concentration was 7.8%.

Factor 4 is distinguished by high levels of O_3 (91.4%) along with 5NSA (80.0%) and 3NSA (86.9%), and is recognized as secondary formation associated with gas-phase reaction (GR). Atmospheric O_3 is the major source of OH radicals, which dominate the secondary formation of NACs from precursors (Harrison et al., 2005). Additionally, field observations and experimental studies have confirmed that NSAs primarily originate from secondary oxidations in the gas phase (Wang et al., 2018; Yuan et al., 2021). The contribution of secondary formation to the NACs concentration was 21.5%.

Text S4. Ensemble machine learning model

Text S4.1. Base models

Random forest (RF) is an ensemble learning technique that constructs multiple decision trees based on bagging theory (Breiman, 2001). RF improves predictive accuracy and controls overfitting by averaging the results of multiple trees, each built from a random subset of the data. This method enhances model robustness, reduces variance, and makes it well-suited for handling large datasets with complex interactions (Requia et al., 2020). Its inherent feature importance evaluation also provides insights into the significance of various predictors (Petkovic et al., 2017).

Extreme gradient boosting (XGBoost), a gradient boosting algorithm, optimizes model performance by sequentially building and combining decision trees. XGBoost incorporates regularization techniques to prevent overfitting and utilizes parallel processing for efficiency, effectively handling large datasets and complex relationships. The XGBoost model has the advantage of superior predictive capabilities and computational efficiency (Fatahi et al., 2022; Gui et al., 2020).

Similar to the XGBoost model, the light gradient boosting machine (LightGBM) is also a gradient boosting technique that leverages tree-based learning algorithms. It utilizes a histogram-based approach for efficient training, significantly reducing computation time and memory usage (Ke et al., 2017). LightGBM handles large datasets and complex features with high accuracy by employing techniques such as gradient-based one-side sampling and exclusive feature bundling. Its advantages include faster training speed, lower memory consumption, and effective handling of categorical features, which collectively enhance predictive performance and scalability (Kang et al., 2021; Ju et al., 2019; Pham et al., 2021).

Multilayer perceptron (MLP) algorithm is a feedforward neural network consisting of an input layer, one or more hidden layers, and an output layer. Each layer is fully connected to the subsequent layer, and MLP uses backpropagation to adjust weights and biases during training. This model can achieve flexibility in modelling intricate data structures, adaptability to various types of tasks, and effectiveness in both regression and classification problems (Reifman and Feldman, 2002; Wang et al., 2023).

The performance of ML approaches is significantly dependent on the hyperparameters, and the optimal values of tuning hyperparameters for the four base learners (RF, XGBoost, LightGBM, and MLP) are listed in Table S2.

Text S4.2. Evaluation index

The coefficient of determination (R^2) evaluates the performance of regression model and quantifies how well the independent variables explain the variability of the dependent variable. R^2 can be calculated according to the Eq. (S2) to (S4) (Spiess and Neumeyer, 2010):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tol}} \tag{S2}$$

$$SS_{res} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad \text{(S3)}$$

$$SS_{tol} = \sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2 \quad \text{(S4)}$$

Root Mean Squared Error (RMSE) measures the square root of the average of the squared differences between the observed actual outcomes and the predictions. Mean Absolute Error (MAE) calculates the average of the absolute differences between the observed actual outcomes and the predictions. Moreover, lower RMSE and MAE values indicate better model performance, and the formulas are as follows (Chai and Draxler, 2014):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad \text{(S5)}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad \text{(S6)}$$

where SS_{res} is the residual sum of squares, SS_{tol} is total sum of squares, y_i and $\hat{y_i}$ are the observed and predicted values, respectively, $\bar{y_i}$ is the mean of observed values, and n is the number of samples.

Text S4.3. SHAP interpretability

Shapley Additive Explanations (SHAP), originating from cooperative game theory (Shapley, 1997), explains the importance of individual features in ML models by evaluating their marginal contributions with SHAP values (Ancona et al., 2019). For each predicted sample, SHAP fairly distributes the contribution values among all features, providing a comprehensive understanding of the relationship between the features and predictions (Hou et al., 2022), as shown in Eq. (S7):

$$f(x) = \varphi_0(f) + \sum_{i=1}^{M} \varphi_i \quad \text{(S7)}$$

where $f(x)$ denotes the predicted value for each sample, $\varphi_0(f)$ is the expected concentration of the model prediction (f) on all samples, M is the number input features, and φ_i is interpreted as Shapley value of *i*-th factor, which represents the contribution of feature *i* and can be expressed as Eq. (8):

$$\varphi_i = \sum_{S\subseteq\{1,2,\ldots\ldots,M\}\backslash\{i\}} \frac{|S|!(M-|S|-1)!}{M!}\left[f(S\cup\{i\}) - f(S)\right] \quad \text{(S8)}$$

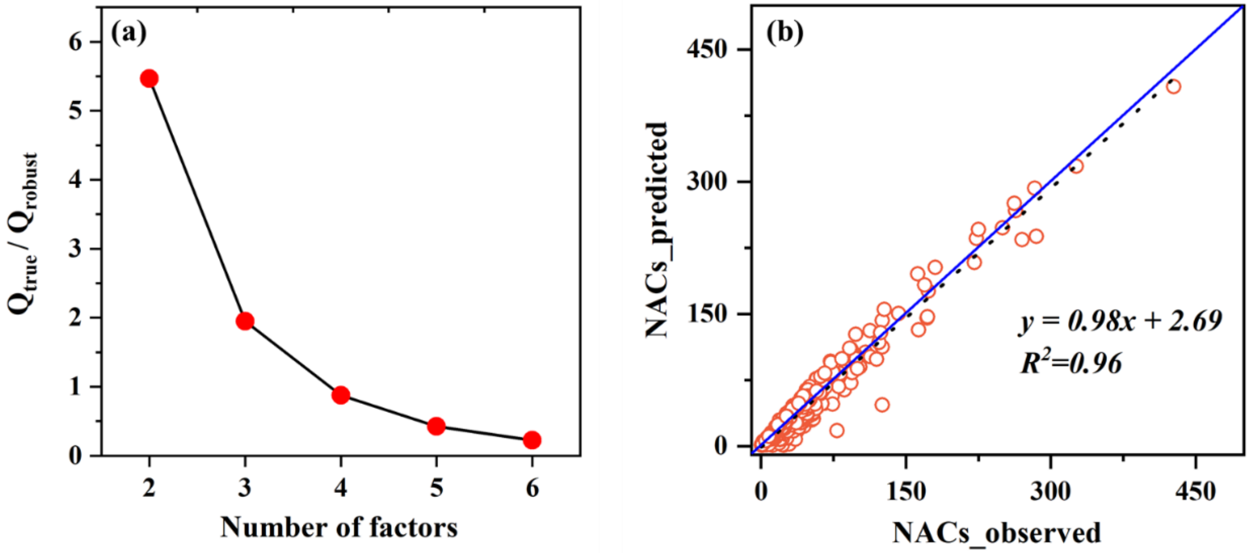
where S is a subset of features excluding feature i, $f(S\cup\{i\})$ is the model prediction when features in subset S and feature i are included, and $f(S)$ is the model prediction when only features in subset S are included.

Table S1. Hyperparameter settings for four base learners.

Model	Hyperparameters	Value
Random forest (RF)	Number of trees	300
	Maximum tree depth	10
	Minimum number of samples required to split an internal node	4
	Minimum number of samples required to be at a leaf node	2
Extreme Gradient Boosting (XGBoost)	Number of trees	300
	Maximum tree depth	3
	Learning rate	0.1
	Subsample	0.8
	Colsample_bytree	1.0
Light Gradient Boosting Machine (LightGBM)	Number of trees	400
	Maximum tree depth	5
	Learning rate	0.1
	Subsample	0.6
	Colsample_bytree	0.6
	Number of leaves	20
Multilayer Perceptron (MLP)	Hidden layer and the number of neurons	1 hidden layer with 100 neurons in each layer
	Activation function	relu
	L2 regularization	10^{-4}
	Tolerance for the optimization	10^{-4}

Table S2. Evaluation index results of NPs, NCs, and NSAs for the EML model.

Compounds	RMSE	MAE	CV-R^2
Nitrophenols (NPs)	5.49	3.13	0.90
Nitrocatechols (NCs)	4.96	2.97	0.85
Nitrosalicylic acids (NSAs)	0.63	0.44	0.93



195 **Figure S1. (a) Q_{true}/Q_{robust} ratios changes with the number of factors and (b) comparison of predicted values by PMF model and observed values of NACs concentrations in PMF analysis.**

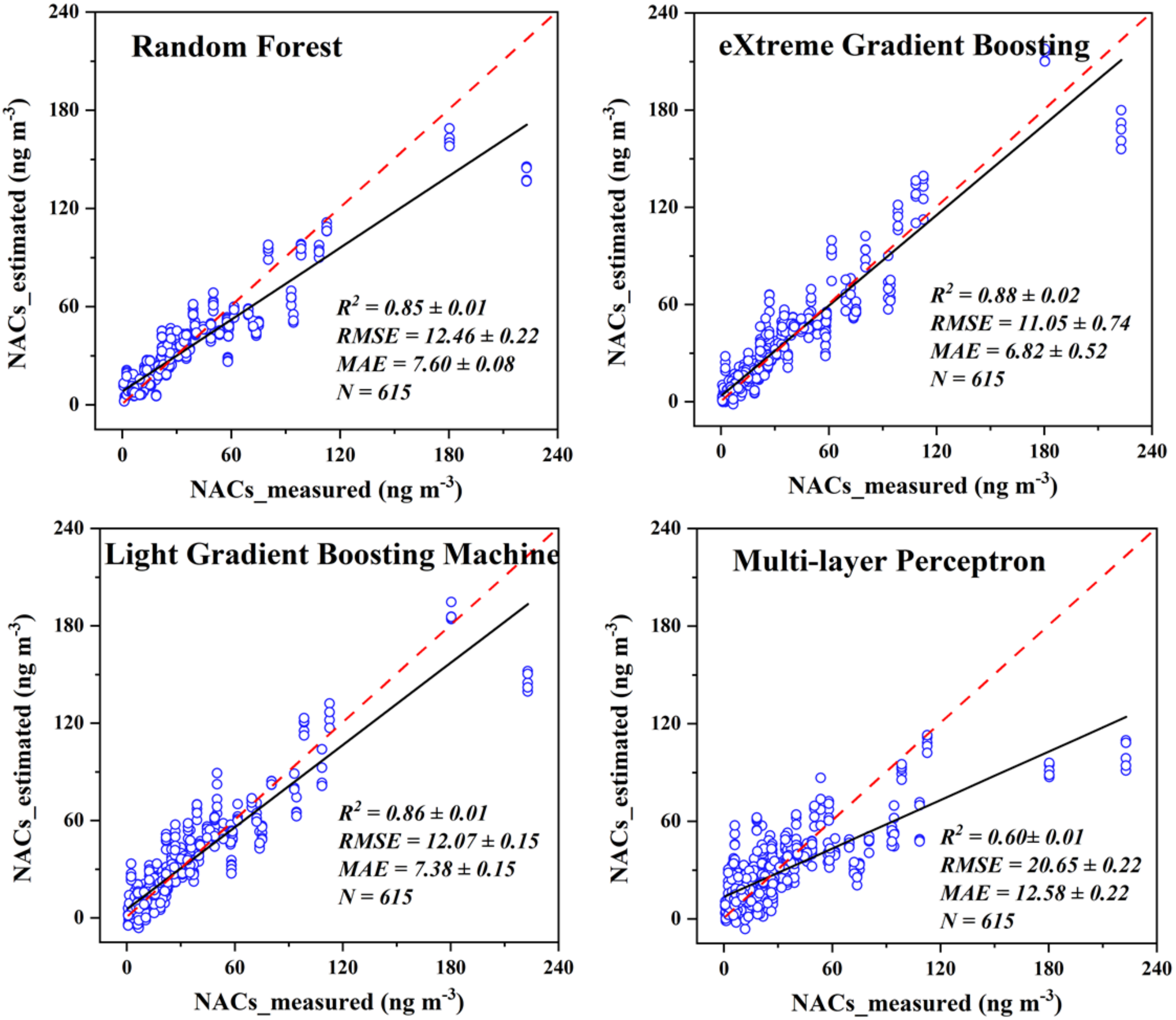


Figure S2. The scatter plots of cross-validation results for simulated and observed NACs on the testing data (obtained after repeating the model five times) by different base models. The red dashed line denotes the best fit line.

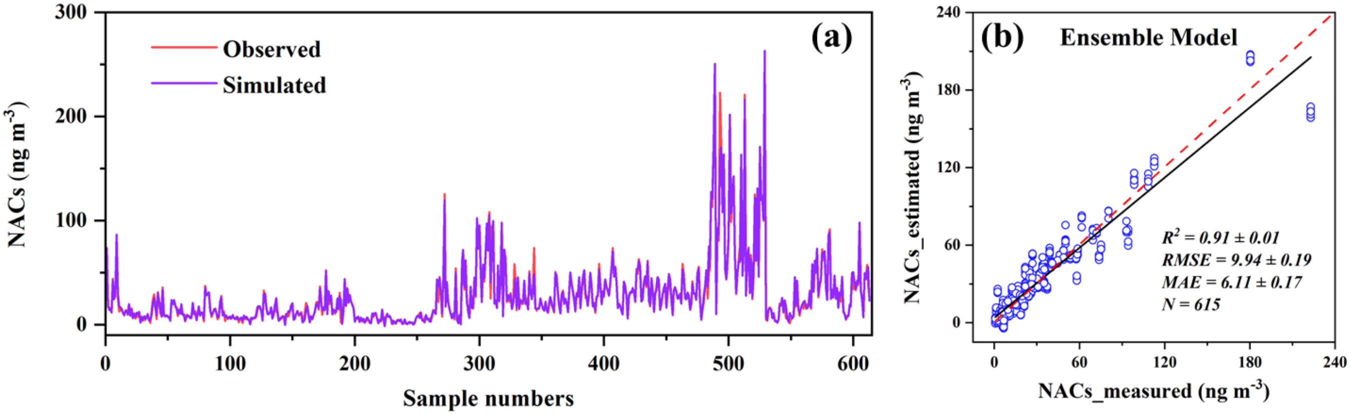


Figure S3. (a) Comparison of EML simulated and observed NACs concentrations for all samples. (b) The scatter plots of cross-validation results for simulated and observed NACs on the testing data (obtained after repeating the model five times) by ensemble machine learning.

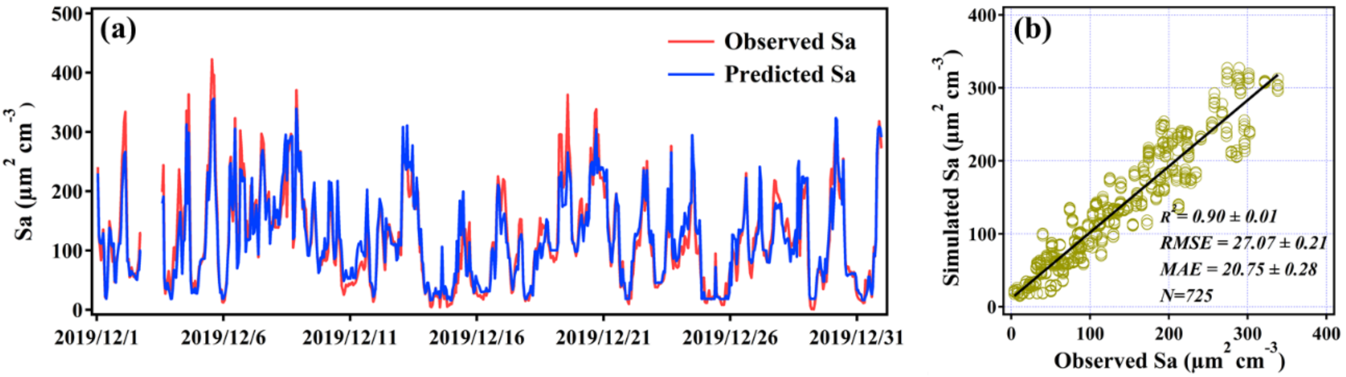


Figure S4. (a) Time series of RF model simulated and observed Sa data during the winter period at Mount Tai. (b) The linear fit between observed and RF model simulated Sa data (obtained after repeating the model five times).

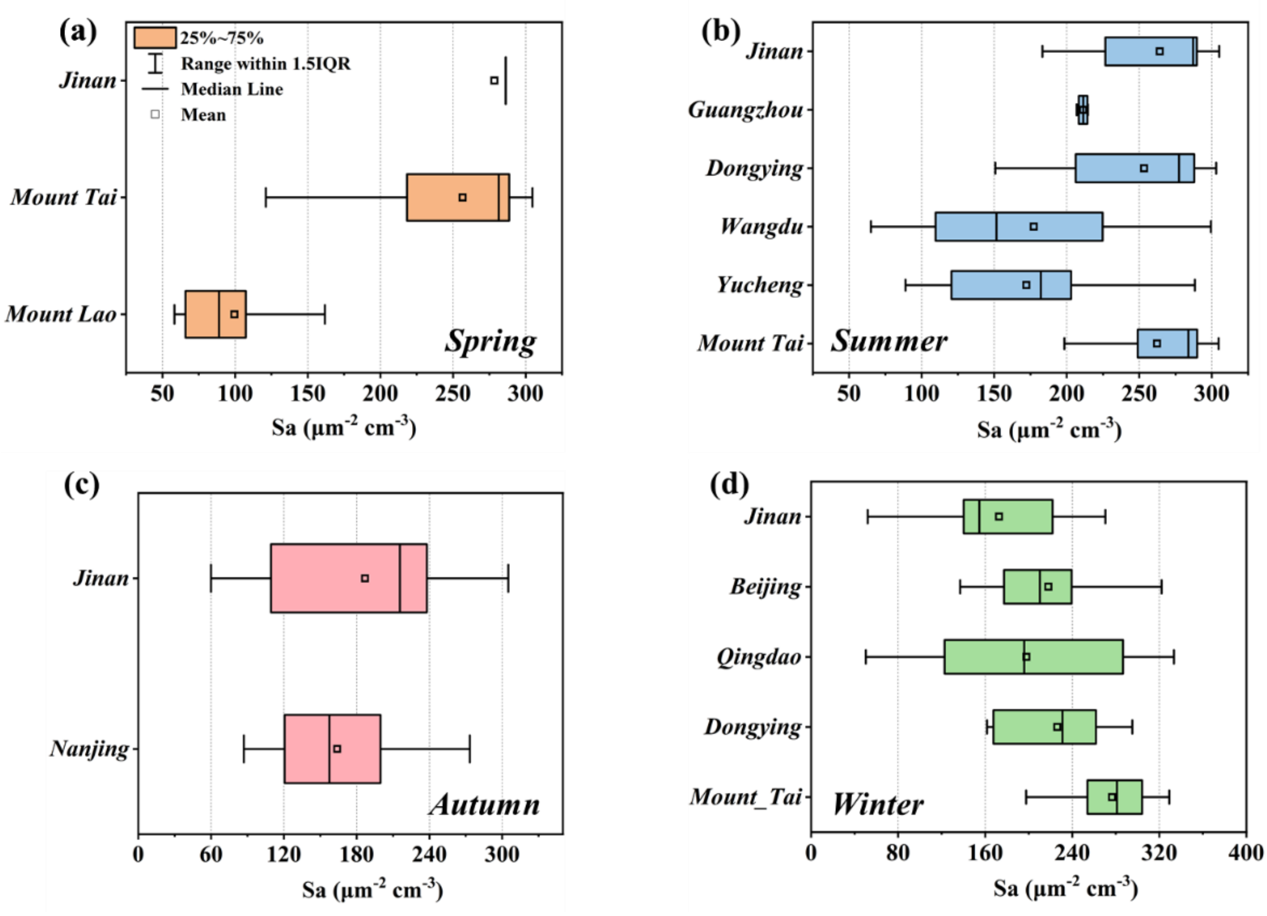
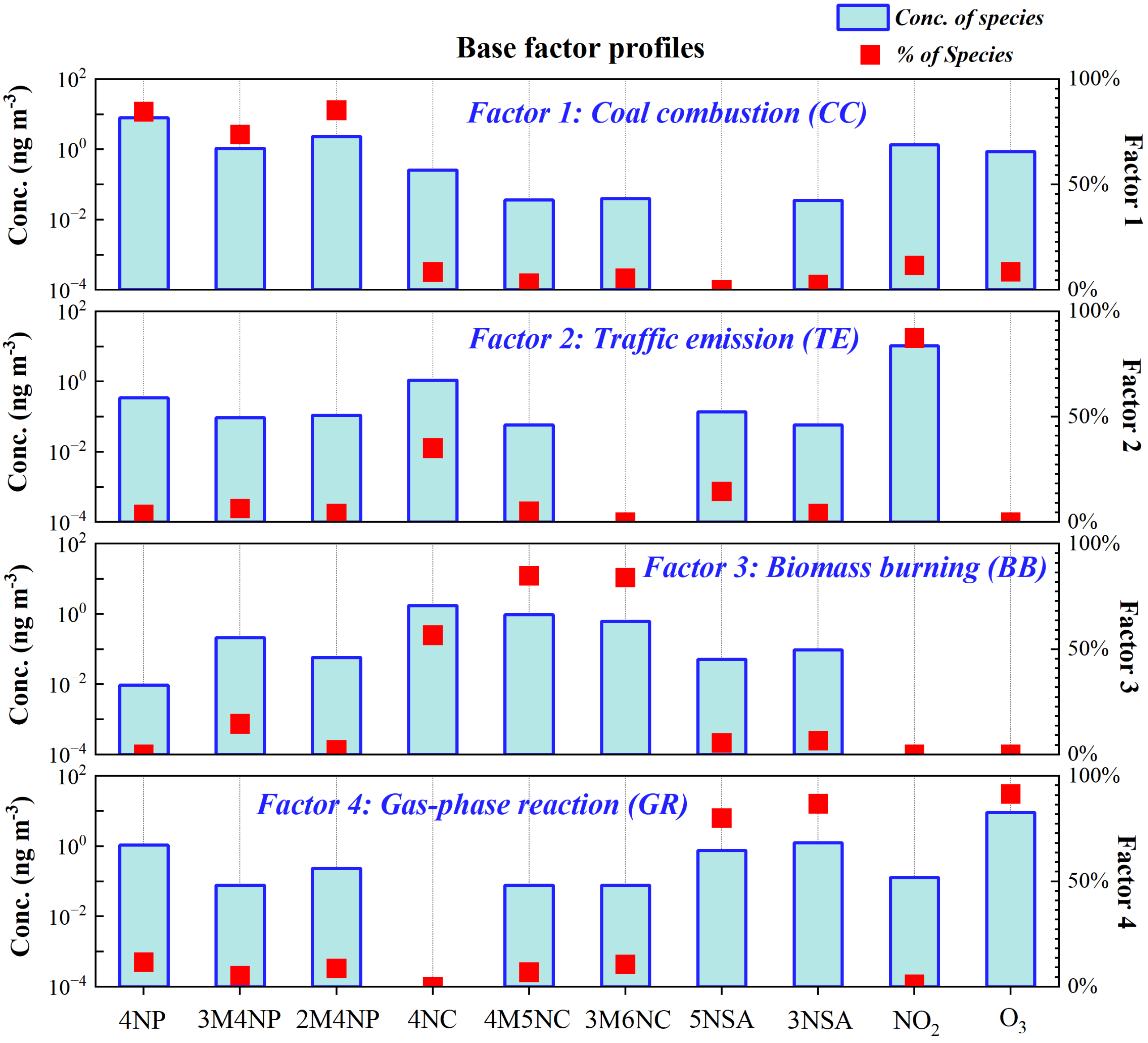


Figure S5. The simulated Sa data based on trained RF model in (a) spring, (b) summer, (c) autumn, and (d) winter at different sampling sites, respectively.



215 **Figure S6. Source profile of resolved factors by PMF model.**

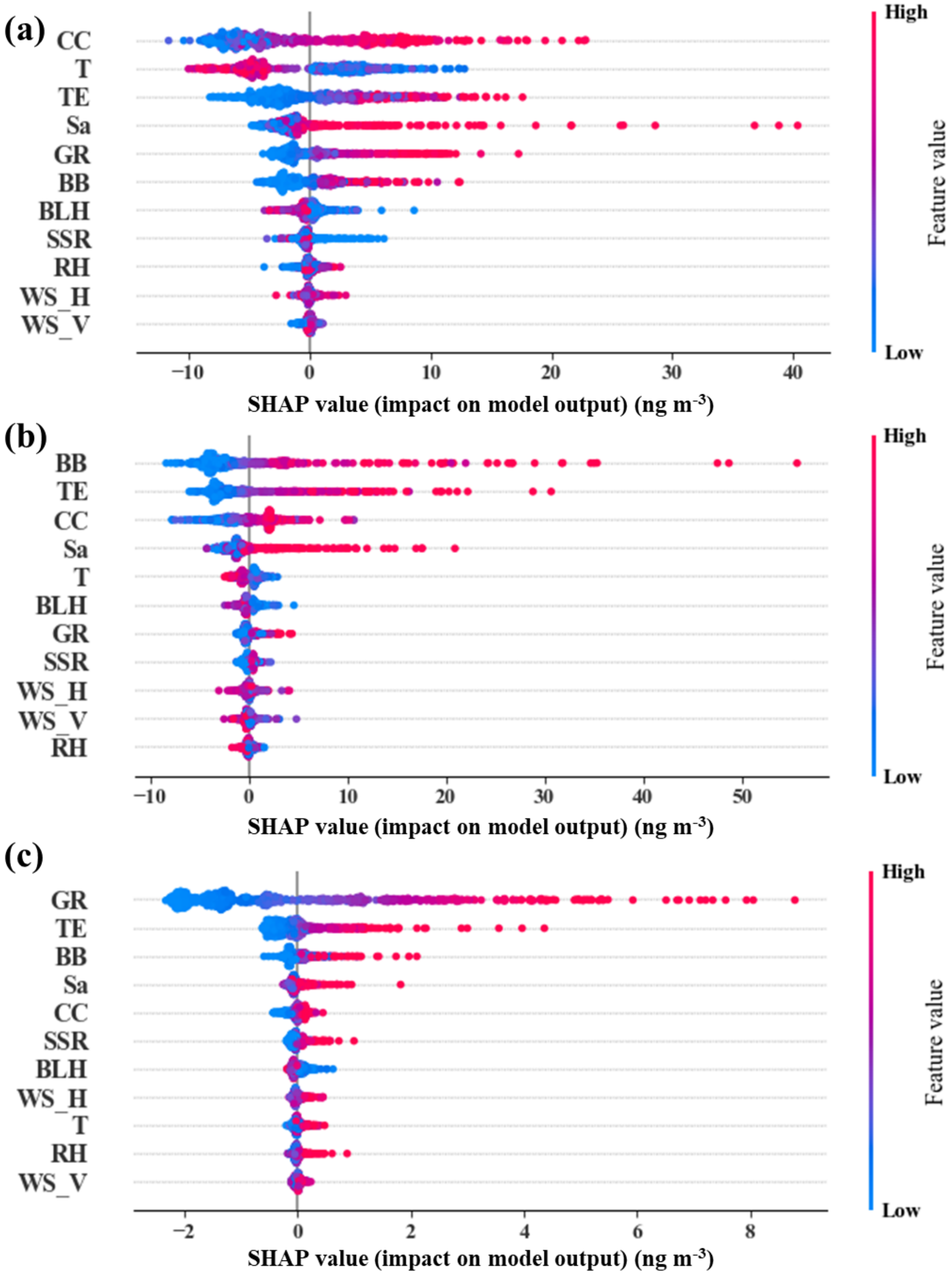
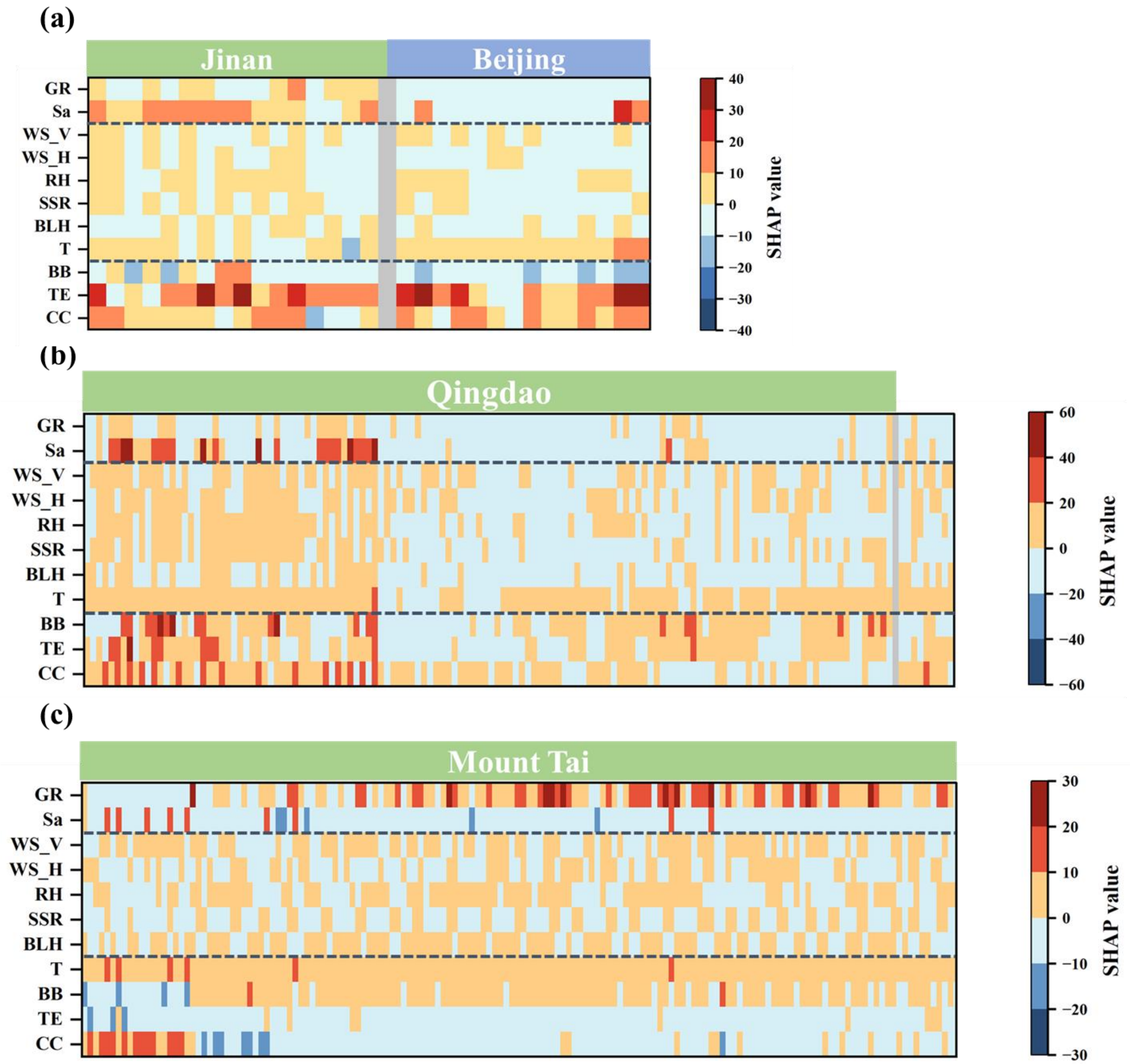


Figure S7. Summary plots of the SHAP interaction matrix values for (a) NPs, (b) NCs, and (c) NSAs, respectively.



220 **Figure S8. Heat maps for the contribution of single factor to each sample in the formation and loss of NACs in the (a) urban, (b) rural, and (c) mountain areas in winter.**

References

Ancona, M., Öztireli, C., and Gross, M.: Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation, 2019.

Bi, X., Lin, Q., Peng, L., Zhang, G., Wang, X., Brechtel, F. J., Chen, D., Li, M., Peng, P. a., Sheng, G., and Zhou, Z.: In situ detection of the chemistry of individual fog droplet residues in the Pearl River Delta region, China, J. Geophys. Res.: Atmos., 121, 9105-9116, https://doi.org/10.1002/2016JD024886, 2016.

Breiman, L.: Random Forests, Machine Learning, 45, 5-32, https://doi.org/10.1023/A:1010933404324, 2001.

Cao, X., Hao, X., Shen, X., Jiang, X., Wu, B., and Yao, Z.: Emission characteristics of polycyclic aromatic hydrocarbons and nitro-polycyclic aromatic hydrocarbons from diesel trucks based on on-road measurements, Atmos. Environ., 148, 190-196, https://doi.org/10.1016/j.atmosenv.2016.10.040, 2017.

Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247-1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.

Claeys, M., Vermeylen, R., Yasmeen, F., Gómez-González, Y., Chi, X., Maenhaut, W., Mészáros, T., and Salma, I.: Chemical characterisation of humic-like substances from urban, rural and tropical biomass burning environments using liquid chromatography with UV/vis photodiode array detection and electrospray ionisation mass spectrometry, Environ. Chem., 9, 273-284, https://doi.org/10.1071/en11163, 2012.

Ding, A. J., Fu, C. B., Yang, X. Q., Sun, J. N., Zheng, L. F., Xie, Y. N., Herrmann, E., Nie, W., Petäjä, T., Kerminen, V. M., and Kulmala, M.: Ozone and fine particle in the western Yangtze River Delta: an overview of 1 yr data at the SORPES station, Atmos. Chem. Phys., 13, 5813-5830, https://doi.org/10.5194/acp-13-5813-2013, 2013.

Fatahi, R., Nasiri, H., Dadfar, E., and Chehreh Chelgani, S.: Modeling of energy consumption factors for an industrial cement vertical roller mill by SHAP-XGBoost: a "conscious lab" approach, Sci. Rep., 12, 7543, https://doi.org/10.1038/s41598-022-11429-9, 2022.

Gui, K., Che, H., Zeng, Z., Wang, Y., Zhai, S., Wang, Z., Luo, M., Zhang, L., Liao, T., Zhao, H., Li, L., Zheng, Y., and Zhang, X.: Construction of a virtual $PM_{2.5}$ observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model, Environ. Int., 141, 105801, https://doi.org/10.1016/j.envint.2020.105801, 2020.

Harrison, M. A. J., Barra, S., Borghesi, D., Vione, D., Arsene, C., and Olariu, R. L.: Nitrated phenols in the atmosphere: a review, Atmos. Environ., 39, 231-248, https://doi.org/10.1016/j.atmosenv.2004.09.044, 2005.

Hong, Y., Cao, F., Fan, M.-Y., Lin, Y.-C., Bao, M., Xue, Y., Wu, J., Yu, M., Wu, X., and Zhang, Y.-L.: Using machine learning to quantify sources of light-absorbing water-soluble humic-like substances (HULISws) in Northeast China, Atmos. Environ., 291, 119371, https://doi.org/10.1016/j.atmosenv.2022.119371, 2022.

Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., and Feng, Y.: Revealing Drivers of Haze Pollution by Explainable Machine Learning, Environ. Sci. Technol. Lett., 9, 112-119, https://doi.org/10.1021/acs.estlett.1c00865, 2022.

Iinuma, Y., Böge, O., Gräfe, R., and Herrmann, H.: Methyl-nitrocatechols: atmospheric tracer compounds for biomass burning secondary organic aerosols, Environ. Sci. Technol., 44, 8453-8459, https://doi.org/10.1021/es102938a, 2010.

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H., and Rehman, M. U.: A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting, IEEE Access, 7, 28309-28318, https://doi.org/10.1109/ACCESS.2019.2901920, 2019.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S.: Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia, Environ. Pollut., 288, 117711, https://doi.org/10.1016/j.envpol.2021.117711, 2021.

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, 3149-3157, https://doi.org/10.5555/3294996.3295074, 2017.

Liu, Z., Li, M., Wang, X., Liang, Y., Jiang, Y., Chen, J., Mu, J., Zhu, Y., Meng, H., Yang, L., Hou, K., Wang, Y., and Xue, L.: Large contributions of anthropogenic sources to amines in fine particles at a coastal area in northern China in winter, Sci. Total Environ., 839, 156281, https://doi.org/10.1016/j.scitotenv.2022.156281, 2022.

Lu, C. Y., Wang, X. F., Li, R., Gu, R. R., Zhang, Y. X., Li, W. J., Gao, R., Chen, B., Xue, L. K., and Wang, W. X.: Emissions of fine particulate nitrated phenols from residential coal combustion in China, Atmos. Environ., 203, 10-17, https://doi.org/10.1016/j.atmosenv.2019.01.047, 2019a.

Lu, C. Y., Wang, X. F., Dong, S. W., Zhang, J., Li, J., Zhao, Y. N., Liang, Y. H., Xue, L. K., Xie, H. J., Zhang, Q. Z., and Wang, W. X.: Emissions of fine particulate nitrated phenols from various on-road vehicles in China, Environ. Res., 179, 108709, https://doi.org/10.1016/j.envres.2019.108709, 2019b.

Petkovic, D., Altman, R., Wong, M., and Vigil, A.: Improving the explainability of Random Forest classifier-user centered approach, in: Biocomputing 2018, World Scientific, 204-215, https://doi.org/10.1142/9789813235533_0019, 2017.

Pham, T. D., Yokoya, N., Nguyen, T. T. T., Le, N. N., Ha, N. T., Xia, J., Takeuchi, W., and Pham, T. D.: Improvement of Mangrove Soil Carbon Stocks Estimation in North Vietnam Using Sentinel-2 Data and Machine Learning Approach, GIScі. Remote Sens., 58, 68-87, https://doi.org/10.1080/15481603.2020.1857623, 2021.

Reifman, J. and Feldman, E. E.: Multilayer perceptron for nonlinear programming, Comput. Oper. Res., 29, 1237-1250, https://doi.org/10.1016/S0305-0548(01)00027-2, 2002.

Ren, Y., Wei, J., Wu, Z., Ji, Y., Bi, F., Gao, R., Wang, X., Wang, G., and Li, H.: Chemical components and source identification of $PM_{2.5}$ in non-heating season in Beijing: The influences of biomass burning and dust, Atmos. Res., 251, 105412, https://doi.org/10.1016/j.atmosres.2020.105412, 2021.

Requia, W. J., Di, Q., Silvern, R., Kelly, J. T., Koutrakis, P., Mickley, L. J., Sulprizio, M. P., Amini, H., Shi, L., and Schwartz, J.: An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-Level Ozone in the Contiguous United States, Environ. Sci. Technol., 54, 11037-11047, https://doi.org/10.1021/acs.est.0c01791, 2020.

Schauer, J. J., Kleeman, M. J., Cass, G. R., and Simoneit, B. R. T.: Measurement of emissions from air pollution sources. 5. C_1-C_{32} organic compounds from gasoline-powered motor vehicles, Environ. Sci. Technol., 36, 1169-1180, https://doi.org/10.1021/es0108077, 2002.

Shapley, L.: Classics in Game Theory 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317, in, edited by: Kuhn, H. W., Princeton University Press, 69-79, https://doi.org/10.1515/9781400829156-012, 1997.

Simoneit, B. R. T., Bi, X., Oros, D. R., Medeiros, P. M., Sheng, G., and Fu, J.: Phenols and Hydroxy-PAHs (Arylphenols) as Tracers for Coal Smoke Particulate Matter: Source Tests and Ambient Aerosol Assessments, Environ. Sci. Technol., 41, 7294-7302, https://doi.org/10.1021/es071072u, 2007.

Spiess, A.-N. and Neumeyer, N.: An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach, BMC Pharmacol., 10, 6, https://doi.org/10.1186/1471-2210-10-6, 2010.

Tham, Y. J., Wang, Z., Li, Q., Yun, H., Wang, W., Wang, X., Xue, L., Lu, K., Ma, N., Bohn, B., Li, X., Kecorius, S., Größ, J., Shao, M., Wiedensohler, A., Zhang, Y., and Wang, T.: Significant concentrations of nitryl chloride sustained in the morning: investigations of the causes and impacts on ozone production in a polluted region of northern China, Atmos. Chem. Phys., 16, 14959-14977, https://doi.org/10.5194/acp-16-14959-2016, 2016.

Tremp, J., Mattrel, P., Fingler, S., and Giger, W.: Phenols and nitrophenols as tropospheric pollutants: Emissions from automobile exhausts and phase transfer in the atmosphere, Water Air Soil Pollut., 68, 113-123, https://doi.org/10.1007/BF00479396, 1993.

Wang, L., Zhao, Y., Shi, J., Ma, J., Liu, X., Han, D., Gao, H., and Huang, T.: Predicting ozone formation in petrochemical industrialized Lanzhou city by interpretable ensemble machine learning, Environ. Pollut., 318, 120798, https://doi.org/10.1016/j.envpol.2022.120798, 2023.

Wang, L., Wang, X., Gu, R., Wang, H., Yao, L., Wen, L., Zhu, F., Wang, W., Xue, L., Yang, L., Lu, K. D., Chen, J., Wang, T., Zhang, Y., and Wang, W.: Observations of fine particulate nitrated phenols in four sites in northern China: concentrations, source apportionment, and secondary formation, Atmos. Chem. Phys., 18, 4349-4359, https://doi.org/10.5194/acp-18-4349-2018, 2018.

Wang, X., Gu, R., Wang, L., Xu, W., Zhang, Y., Chen, B., Li, W., Xue, L., Chen, J., and Wang, W.: Emissions of fine particulate nitrated phenols from the burning of five common types of biomass, Environ. Pollut., 230, 405-412, https://doi.org/10.1016/j.envpol.2017.06.072, 2017a.

Wang, X., Wang, H., Xue, L., Wang, T., Wang, L., Gu, R., Wang, W., Tham, Y. J., Wang, Z., and Yang, L.: Observations of N_2O_5 and $ClNO_2$ at a polluted urban surface site in North China: High N_2O_5 uptake coefficients and low $ClNO_2$ product yields, Atmos. Environ., 156, 125-134, https://doi.org/10.1016/j.atmosenv.2017.02.035, 2017b.

Wang, Z., Wang, W., Tham, Y. J., Li, Q., Wang, H., Wen, L., Wang, X., and Wang, T.: Fast heterogeneous N_2O_5 uptake and $ClNO_2$ production in power plant and industrial plumes observed in the nocturnal residual layer over the North China Plain, Atmos. Chem. Phys., 17, 12361-12378, https://doi.org/10.5194/acp-17-12361-2017, 2017c.

Wu, X., Cao, F., Haque, M., Fan, M.-Y., Zhang, S.-C., and Zhang, Y.-L.: Molecular composition and source apportionment of fine organic aerosols in Northeast China, Atmos. Environ., 239, 117722, https://doi.org/10.1016/j.atmosenv.2020.117722, 2020.

Yao, L., Yang, L., Chen, J., Wang, X., Xue, L., Li, W., Sui, X., Wen, L., Chi, J., Zhu, Y., Zhang, J., Xu, C., Zhu, T., and Wang, W.: Characteristics of carbonaceous aerosols: Impact of biomass burning and secondary formation in summertime in a rural area of the North China Plain, Sci. Total Environ., 557-558, 520-530, https://doi.org/10.1016/j.scitotenv.2016.03.111, 2016.

Yuan, W., Huang, R., Yang, L., Wang, T., Duan, J., Guo, J., Ni, H., Chen, Y., Chen, Q., Li, Y., Dusek, U., O'Dowd, C., and Hoffmann, T.: Measurement report: $PM_{2.5}$-bound nitrated aromatic compounds in Xi'an, Northwest China – seasonal variations and contributions to optical properties of brown carbon, Atmos. Chem. Phys., 21, 3685-3697, https://doi.org/10.5194/acp-21-3685-2021, 2021.

Zhang, Q., Gao, R., Xu, F., Zhou, Q., Jiang, G., Wang, T., Chen, J., Hu, J., Jiang, W., and Wang, W.: Role of Water Molecule in the Gas-Phase Formation Process of Nitrated Polycyclic Aromatic Hydrocarbons in the Atmosphere: A Computational Study, Environ. Sci. Technol., 48, 5051-5057, https://doi.org/10.1021/es500453g, 2014.

Zhang, Y., Sun, J., Zheng, P., Chen, T., Liu, Y., Han, G., Simpson, I. J., Wang, X., Blake, D. R., Li, Z., Yang, X., Qi, Y., Wang, Q., Wang, W., and Xue, L.: Observations of C_1-C_5 alkyl nitrates in the Yellow River Delta, northern China: Effects of biomass burning and oil field emissions, Sci. Total Environ., 656, 129-139, https://doi.org/10.1016/j.scitotenv.2018.11.208, 2019.