Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jes

JES JOURNAL OF ENVIRONMENTAL SCIENCES

www.jesc.ac.cn

Machine learning and theoretical analysis release the non-linear relationship among ozone, secondary organic aerosol and volatile organic compounds ☆

Feng Wang [1,2], *Zhongcheng Zhang* [1,2], *Gen Wang* [3], *Zhenyu Wang* [1,2], *Mei Li* [4,5], *Weiqing Liang* [1,2], *Jie Gao* [1,2], *Wei Wang* [6,*], *Da Chen* [7], *Yinchang Feng* [1,2], *Guoliang Shi* [1,2,*]

[1] *State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, Tianjin Key Laboratory of Urban Transport Emission Research, College of Environmental Science and Engineering, Nankai University, Tianjin 300350, China*

[2] *CMA-NKU Cooperative Laboratory for Atmospheric Environment-Health Research (CLAER), College of Environmental Science and Engineering, Nankai University, Tianjin 300350, China*

[3] *State Key Laboratory on Odor Pollution Control, Tianjin Academy of Environmental Sciences, Tianjin 300191, China*

[4] *Guangdong Provincial Engineering Research Center for On-line Source Apportionment System of Air Pollution Jinan University, Institute of Mass Spectrometry and Atmospheric Environment, Guangzhou 510632, China*

[5] *Guangdong-Hongkong-Macau Joint Laboratory of Collaborative Innovation for Environmental Quality, Guangzhou 510632, China*

[6] *Trusted AI System Laboratory, College of Computer Science, Nankai University, Tianjin 300350, China*

[7] *Key Laboratory of Civil Aviation Thermal Hazards Prevention and Emergency Response, Civil Aviation University of China, Tianjin 300300, China*

ARTICLE INFO

Article history:
Received 28 May 2021
Revised 22 July 2021
Accepted 22 July 2021
Available online 14 January 2022

Keywords:
VOCs
Machine learning
Photochemical consumption
Ozone formation potential
Secondary organic aerosol formation potential

ABSTRACT

Fine particulate matter ($PM_{2.5}$) and ozone (O_3) pollutions are prevalent air quality issues in China. Volatile organic compounds (VOCs) have significant impact on the formation of O_3 and secondary organic aerosols (SOA) contributing $PM_{2.5}$. Herein, we investigated 54 VOCs, O_3 and SOA in Tianjin from June 2017 to May 2019 to explore the non-linear relationship among O_3, SOA and VOCs. The monthly patterns of VOCs and SOA concentrations were characterized by peak values during October to March and reached a minimum from April to September, but the observed O_3 was exactly the opposite. Machine learning methods resolved the importance of individual VOCs on O_3 and SOA that alkenes (mainly ethylene, propylene, and isoprene) have the highest importance to O_3 formation; alkanes (C_n, $n \geq 6$) and aromatics were the main source of SOA formation. Machine learning methods revealed and emphasized the importance of photochemical consumptions of VOCs to O_3 and SOA formation. Ozone formation potential (OFP) and secondary organic aerosol formation potential (SOAFP) calculated by consumed VOCs quantitatively indicated that more than 80% of the consumed VOCs were alkenes which dominated the O_3 formation, and the impor-

☆ This article is dedicated to Professor Dianxun Wang.

* Corresponding authors.
E-mails: kevinwangwei@nankai.edu.cn (W. Wang), nksgl@nankai.edu.cn (G. Shi).

https://doi.org/10.1016/j.jes.2021.07.026

tance of consumed aromatics and alkenes to SOAFP were 40.84% and 56.65%, respectively. Therein, isoprene contributed the most to OFP at 41.45% regardless of the season, while aromatics (58.27%) contributed the most to SOAFP in winter. Collectively, our findings can provide scientific evidence on policymaking for VOCs controls on seasonal scales to achieve effective reduction in both SOA and O_3.

Introduction

Fine particulate matter ($PM_{2.5}$) and ozone (O_3) are the main air-pollutants that pose substantial adverse impacts on climate and human health (Colmer et al., 2020; Lu et al., 2018; Wang et al., 2020b; Zhang et al., 2019a; Zheng et al., 2019). Secondary organic aerosol (SOA) is a major component of $PM_{2.5}$, even accounting for about half in winter haze-pollution (An et al., 2019; Peng et al., 2021; Wang et al., 2013b). SOA formation is controlled by two processes, namely, the oxidation of volatile organic compounds (VOCs) and the semi-volatile VOCs (SVOCs) gas-to-particle conversion (Wang et al., 2021). Whereas O_3 is produced by photochemical oxidation of VOCs catalysed by hydroxyl radicals (OH radical) (Meng et al., 1997; Tan et al., 2018; Vermeuel et al., 2019). The aforementioned complex mechanism exerts a direct influence on O_3 and SOA formation that makes a non-linear relationship towards VOCs (Buysse et al., 2019; Chen et al., 2019; Wu et al., 2020). Recently, VOC emissions increased by 2% from 2013 to 2017 in China (Li et al., 2019), which makes VOCs research and control are prioritized. Extensive research from field observations and model simulations are conducted on detailed VOCs chemical and physical processes leading to O_3 and SOA pollution (Niu et al., 2016; Shang et al., 2020; Tan et al., 2018; Yang et al., 2019). For example, 20%–40% reductions of VOCs emission are effective in reducing the maximum daily average 8 hr (MDA8) O_3 by 5–10 ppbv (Kang et al., 2021; Li et al., 2019). While, due to the inaccuracy of meteorological conditions, precursor emissions, and complex chemical processes (Li et al., 2021; Wang et al., 2021), less is known about the SOA concentration in response to VOCs. Therefore, the non-linear response of O_3 and SOA to VOCs concentrations changing remains uncertain.

Machine learning method is an emerging concern driven by data changing, and can effectively establish the complex and non-linear relationships and interactions between input features and output predictor variables. It has been increasingly applied to predict air pollutant (PM_{10}, $PM_{2.5}$, NO_x, and NH_3 et.al.) concentrations (Choubin et al., 2020; Li et al., 2020c; Zhang et al., 2019b) and to analyze the variable importance for different pollution episodes (Gao et al., 2020), which has a better performance compared to traditional methods. Therein, random forest (RF) model outperformed other methods (Boosted Regression Trees, Support Vector Machine) with higher cross validation (CR) values (higher model accuracy) for $PM_{2.5}$ and NO_x prediction (Choubin et al., 2020). However, current machine learning methods mostly focused on prediction accuracy, yet comparative individual interpreting predictions (Lundberg and Lee, 2017). Shapley additive explanation regression (SHAP) value reveals how the input variable values impacts the corresponding predictions of every sample, which has been performed on medical problems (Lundberg et al., 2020), single-cell classification (Xie et al., 2020). Therefore, the data driven machine learning method can better characterize and interprete the non-linear relationship between ozone and SOA on VOCs.

Herein, the two years continuous measurement of VOCs was carried out in this study. The relationship of O_3 and SOA to VOCs concentrations was investigated based on the observed data. Machine learning method was applied to explore the above non-linear response. Photochemical loss of VOCs was estimated to explore the accordingly O_3 and SOA formation. Furthermore, the importance of consumed VOCs to ozone formation potential (OFP) and secondary organic aerosol formation potential (SOAFP) were studied, and their seasonal variation was analyzed. Machine learning methods and theoretical analysis present comparable results and provide insight on effective mitigation policies for O_3 and SOA control.

1. Materials and methods

1.1. *Sampling and monitoring*

The sampling site is conducted at Nankai University Air Quality Research Supersite (NKAQRS) (38°59′N, 117°19′E) in Tianjin, from June 2017 to May 2019. It is a typical rural area, where is far from major highways and high traffic zones. Detailed sampling site information is available in previous reports (Dai et al., 2020; Gu et al., 2020; Tian et al., 2020). Hourly concentration of $PM_{2.5}$, O_3, fifty-four VOCs species, organic carbon (OC) and element carbon (EC) are measured by various instruments. Detailed description is in Appendix A Section S1. Given the severe photochemical reaction that makes the assumption of VOCs and the generation of O_3, in this work, we utilize daytime measured data from 8:00 to 19:00 local time (LT), when is typical of O_3 formation period (Zou et al., 2015).

1.2. *Machine learning methods*

The random forest (RF) model, a novel supervised machine learning algorithm, uses Decision Trees for data classification and have been widely applied for analyzing the input variable importance in different air pollution episodes (Gao et al., 2020). The variable importance (%) represents the promoting ability of various VOCs to O_3 and SOA generation. In this study, firstly, fifty-four VOCs species data are as input variables to predict O_3. And twenty-eight VOCs species data are as input variables to SOA, due to partly of VOCs (alkanes (C_n,

$n \geq 6$), isoprene and aromatics) are attributed to SOA formation. Secondly, entire O_3 and SOA dataset are respectively divided into: training dataset, 80% of the randomly dataset, is used to build the RF model; test dataset, 20% of the randomly dataset, is used to test the prediction. Finally, the variable importance (%) of VOCs species are obtained by random forest model. The latest Python-packages (scikit-learn package, https://scikit-learn.org/stable/) are used to achieve the RF model. As the relative importance computed by RF model is lacking of interpretability, individualized importance of VOCs species are performed using shapley additive explanation regression (SHAP) values (Lundberg et al., 2020; Lundberg and Lee, 2017). In our study, three types of simulation was performed by SHAP value, based on SOA/O_3 radios to distinguish VOCs tending to generate O_3 or SOA. Type 1 represents the higher concentration of SOA with lower concentration of O_3; Type 2 represents the moderate concentration of SOA and O_3; Type 3 represents the higher concentration of O_3 with lower concentration of SOA. The detailed information of SHAP values used in this study and the calculated process are described in Appendix A Section S2.

1.3. *Theoretical analysis methods*

In the ambient environment, the estimation of SOA is highly uncertain due to the strong photochemical loss of VOCs. Various of methodologies can be used to estimate secondary organic aerosol, including EC-tracer method, PMF-based approaches and SOA-tracer method (Srivastava et al., 2021). In this section, we applied the EC-tracer method to estimate SOA. The photochemical consumption of VOCs is calculated based on the assumption that the loss of VOCs in LT is dominated by the oxidation of OH radicals. The OFP and SOAFP are further calculate to explore the importance of consumsed VOCs to O_3 and SOA.

1.3.1. *Total SOA estimation by the EC-tracer method*

EC-tracer method is conducted on the total SOA concentration measurement. It roughly estimates the OC partition onto primary and secondary fractions. By assuming EC and primary OC own the same origins, EC-tracer method estimates secondary organic (SOC) from the magnitude of OC-to-EC ratios (Gao et al., 2019):

$$SOC = OC - EC \times \left(\frac{OC}{EC}\right)_{min} \quad (1)$$

$$SOA = SOC \times 1.8 \quad (2)$$

where the OC/EC ratio was estimated using the observed minimum OC/EC ratio.

1.3.2. *Photochemical consumption of VOCs*

The photochemical consumed VOCs ($VOC_{consumed}$) is defined as their reactions with OH radicals in daytime. With larger OH reaction rate constant, the photochemical consumed VOCs is more significant for O_3 and SOA formation. The photochemical consumed VOCs of the i th species ($VOC_{i,consumed}$) is calculated by Eq. (3) (Gao et al., 2019; Min et al., 2011; Zhang et al., 2020b):

$$VOC_{i,consumed} = VOC_{i,t} \times (\exp(k_i[OH]\Delta t) - 1) \quad (3)$$

where k_i is the reaction rate constant of VOC_i with OH radicals, [OH] is the concentration of OH radicals, Δt is the photochemical age and is given by Eq. (4):

$$\Delta t = \frac{1}{[OH](k_E - k_X)} \times \left(\ln \frac{[E]}{[X]}|t_0 - \ln \frac{[E]}{[X]}|t\right) \quad (4)$$

where k_E and k_X are the reaction rate constants of ethylbenzene (E) and m- and p-xylene (X) with OH radicals, 7.0×10^{-12} and 18.9×10^{-12} cm^3/(molecule•sec), respectively (Gao et al., 2019). $([E]/[X])_t$ is the measured ratio of ethylbenzene to m- and p-xylene. $([E]/[X])_{t0}$ is assumed as the initial mixing ratios of ethylbenzene to m- and p-xylene in the VOCs emission sources (Wang et al., 2013a). Previous studies reported that $([E]/[X])_{t0}$ is related to different sources, such as: vehicular emission and coal combustion (0.3-0.4) (Liu et al., 2008), solvent use (0.4-0.5) (Wang et al., 2013a), and biomass burning (about 0.65) (Liu et al., 2008). This ratio is also used as an indicator for photochemical reactions (Gao et al., 2019) and can be estimated by the least ratio observed in daytime. Therefore, we choose the value of $([E]/[X])_{t0}$ as 0.66 in our study.

1.4. *OFP and SOAFP calculation*

OFP is a parameter to evaluate the impacts of VOCs on O_3 production, which involves two characteristics: kinetic reactivity and mechanism reactivity (Zou et al., 2015). Here, we used the mechanism reactivity of VOCs, the MIR (maximum incremental reactivity)-weighted concentration, to estimate the maximum O_3 concentration variation. OFP was calculated with respect to the VOCs concentration and corresponding MIR by Eq. (5) (Cardelino and Chameides, 1995; Carter, 2010):

$$OFP_i = [VOCs]_i \times MIR_i \quad (5)$$

where OFP_i is the OFP of the ith VOCs species, $[VOCs]_i$ is the concentration of the ith VOCs species, and MIR_i is the MIR for the ith VOCs species. The MIR constant of each VOCs species are updated versions of the research results of Carter (2010) in Table S2.

SOAFP is widely used to measure the VOCs importance to SOA generation (Song et al., 2018). Fractional aerosol coefficient (FAC) method, developed by Grosjean and Seinfeld (1989) and Grosjean (1992), reflects the fraction of SOA resulting from the reaction of VOCs. SOAFP equation is expressed as Eq. (6) (Li et al., 2020a; Song et al., 2018):

$$SOAFP_i = [VOCs]_i \times FAC_i \quad (6)$$

where $SOAFP_i$, $[VOCs]_i$ and FAC_i represent the corresponding SOAFP, VOCs concentration, and fractional aerosol coefficient of certain substance i in VOCs, respectively. The FAC constant of each VOCs species are shown in Table S1.

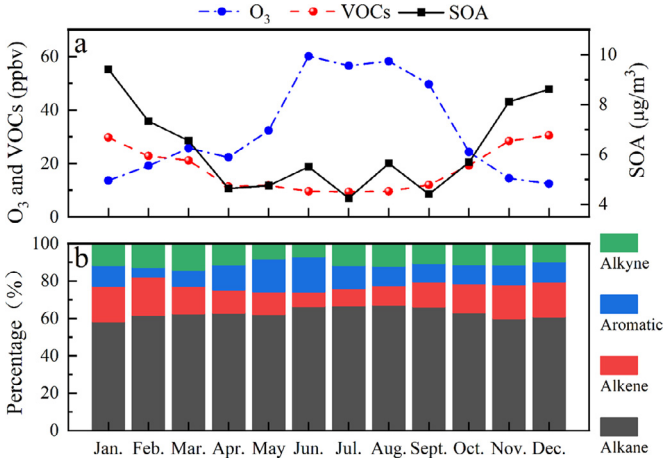

Fig. 1 – (a) Monthly variation concentrations of observed hourly concentration of volatile organic compounds (VOCs), ozone (O_3) and calculated secondary organic aerosol (SOA). Measured local daytime is from 8:00 to 19:00. (b) Disparities in proportions of four classes of VOCs in monthly variations.

2. Results and discussion

2.1. *VOCs, O_3 and SOA concentrations*

Fig. 1a shows the monthly variation concentration characteristics of observed VOCs, O_3, and calculated SOA. VOCs and SOA concentrations are showing a certain regularity variation, being lower from April to September and higher from October to March. Whereas, O_3 concentration has readily apparent opposite variation being higher in summer and lower in winter, which is a typical annual observation variable (Yang et al., 2019). This phenomenon implies that VOCs are negatively correlated with O_3 while positively consistent with SOA. Disparities in VOCs concentration are frequently observed to be correlated with O_3 and SOA pollution situation (Li et al., 2020a).

We go on to more specifically attribute to major VOC species (Fig. 1b). VOCs proportions at any given month is unequally distributed. Alkanes are the most abundant VOCs group, especially in summer, accounting for 66.79% of the total VOCs concentrations. Pervasive studies suggest that alkanes play a significant but not dominant role in O_3 formation (Liu et al., 2021; Song et al., 2018), and this is mainly due to the lower reaction reactivity between alkanes and OH radical (Gligorovski et al., 2018; Zou et al., 2015). Alkenes account for the smallest proportion (9.11%) in summer with highest O_3 concentration (58.33 ppbv). However, alkenes are expected to be tightly correlation with O_3 formation (Gligorovski et al., 2018; Zhang et al., 2020a). Meanwhile, aromatics are abundant in summer contributing about 13.94%. However, as the major aromatics (Table S1), toluene and benzene have been suggested to be favored SOA formation (Gao et al., 2019; Wang et al., 2020a). The proportion of alkyne showed no significant difference. Similar observations made in several studies that alkanes were the most dominant species, followed by aromatics, alkenes and acetylene (Gu et al., 2019; Li et al., 2020b; Zhang et al., 2021). Detailed description for the absolute concentrations of total VOCs is shown in Table S1 and discussed in Appendix A Section S3. Significant difference existed in observed O_3 and SOA concentrations against VOCs. The nonlinear response of O_3 and SOA to VOCs cannot be simply explained by observation data, which we will elaborate on this in the next section.

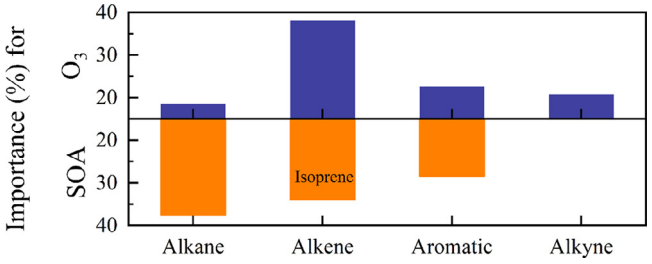

Fig. 2 – Variable importance values of alkanes, alkenes, aromatics, and alkyne, to O_3 and SOA formation, respectively.

2.2. *VOCs species importance estimation by machine learning method*

To quantify the VOCs importances to O_3 and SOA formation, a machine learning method called random forest (RF) model was utilized to calculate the VOCs species importance. Fig. 2 shows the variable importance of VOCs species to O_3 and SOA formation. The cross validation (CV) of the RF results is 0.78 and 0.56, respectively, implying the results can be accepted. As can be seen in Fig. 2, alkenes have the highest importance to O_3 formation (38%), followed by aromatics (23%), alkyne (21%) and alkanes (18%). Whereas for SOA formation, data-driven analysis results show that alkanes (C_n, $n \geq 6$) got predominated occupation (38%) of variable importance, indicating that higher relatively molecular weight substances are the main source of SOA formation due to the rapidly reaction into condensed phase in the atmosphere, which is consistent with the previous studies (Li et al., 2020a). Alkene (isoprene) is of approximately equal importance (34%) to SOA formation. Gen-

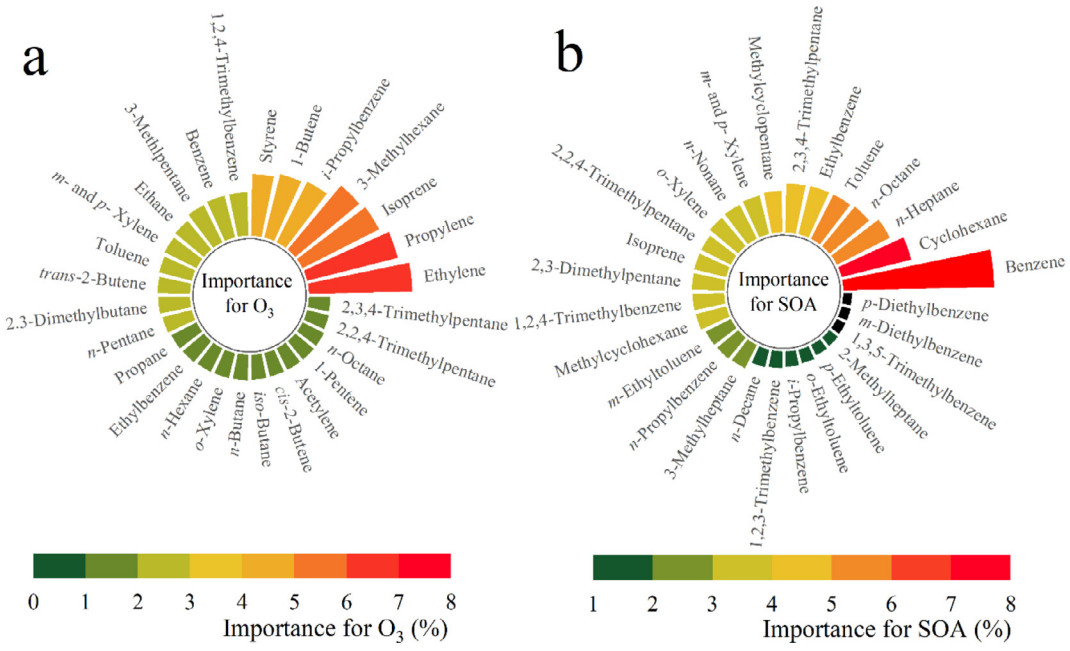

Fig. 3 – Variable importance values of individual VOCs variable importance to (a) O_3 and (b) SOA formation, respectively.

erally, isoprene can react with OH radical and O_3 to produce organic peroxy radical and stabilized Criegee intermediates, respectively (Huang et al., 2008; Peng et al., 2021). Both of them are the important SOA formation pathway. Aromatics showed relatively low but not negligible value (28%) of variable importance. The aforementioned data-driven analysis is specifically consistent with previous studies (Li et al., 2020a; Song et al., 2018).

We further estimated the individual VOCs variable importance to O_3 and SOA formation, respectively, shown in Fig. 3. A good performance of the cross validation (CV, R^2) (0.78 and 0.56 for O_3 and SOA) could reflect the actual applicability of the machine learning method. To O_3 formation, ethylene (6.93%), propylene (6.27%), and isoprene (5.75%) are dominant, indicating that alkenes are the most important variable for O_3 formation which is consistent with the above analysis. To SOA formation, benzene (15.17%), toluene (5.01%), ethylbenzene (4.98%), and *m*- and *p*-xylene (3.99%) accounting for 29.14%, while cyclohexane (7.22%), *n*-heptane (5.98%), and *n*-octane (5.18%) are contributed about 17.92%. This again implies that aromatics are favoured SOA formation and even contribute over 95% to SOAFP (Liu et al., 2021; Qin et al., 2021). Collectively, data-driven analysis tells us the opposite of the observation results that alkenes are important in driving ozone formation, especially ethylene, propylene and isoprene, and aromatics and alkanes (C_n, $n \geq 6$) facilitate SOA formation.

To further identify which variable affects both O_3 and SOA generation simultaneously, SHAP are commonly used to conduct the variable importance. Fig. 4 shows the mean variable importance for three Types, ordered from high to low, to further separate individual VOCs influences on O_3 and SOA. Isoprene has the highest impact, followed by 1,2,4-trimethylbenzene, benzene, 2,3,4-trimethylpentane, *m*- and *p*-xylene and cyclohexane. This means that, to a large extent, isoprene, 1,2,4-trimethylbenzene and benzene are both related to O_3 and SOA. To further explore how the expected output of the importance changes as the VOCs varies, the importance of input 28 VOCs are plotted as a function of variable values. We present the top four key VOCs for Type 1 (high SOA concentration) and Type 3 (high O_3 concentration) with the magnitude of the importance varying by species (Fig. 5). The SHAP value of a variable is supposed to be the *y*-axis and the value of the variable is the *x*-axis. Model taking into account all possible input variable combinations to obtain the mean prediction by adding SHAP values to the base value. Therefore, SHAP values can be negative, which reduce the prediction below the base value. Whereas, the positive SHAP value indicates that this variable is expected to have distinct influence on the model prediction. As shown in Fig. 5, the impact of VOCs concentrations on simulated SHAP values is highly non-linear. The 1,2,4-trimethylbenzene, *m*- and *p*-xylene and benzene show similar variable trends that increased positive importances (Fig. 5a, c and d) and negative importances (Fig. 5f, g and h) of model outcomes with elevated SOA concentrations. This implies that higher concentrations of 1,2,4-trimethylbenzene, *m*- and *p*-xylene and benzene can promote SOA formation. However, isoprene presents contrasting trend that high importances of isoprene are expected to lead to high concentrations of O_3 in Type 3 (Fig. 5e). Other component of VOCs, ranked by their relative importance, are presented in Figs. S1 and S2.

Of particular confusion, observation and machine learning method yield contradicting conclusions. Observational evidence suggested that alkanes and aromatics variation trends is roughly consistent with O_3 and alkenes are corresponding to SOA, whereas machine learning method found that alkenes dominate the production of O_3 and aromatics contribute greatly to the generation of SOA. Generally, SOA for-

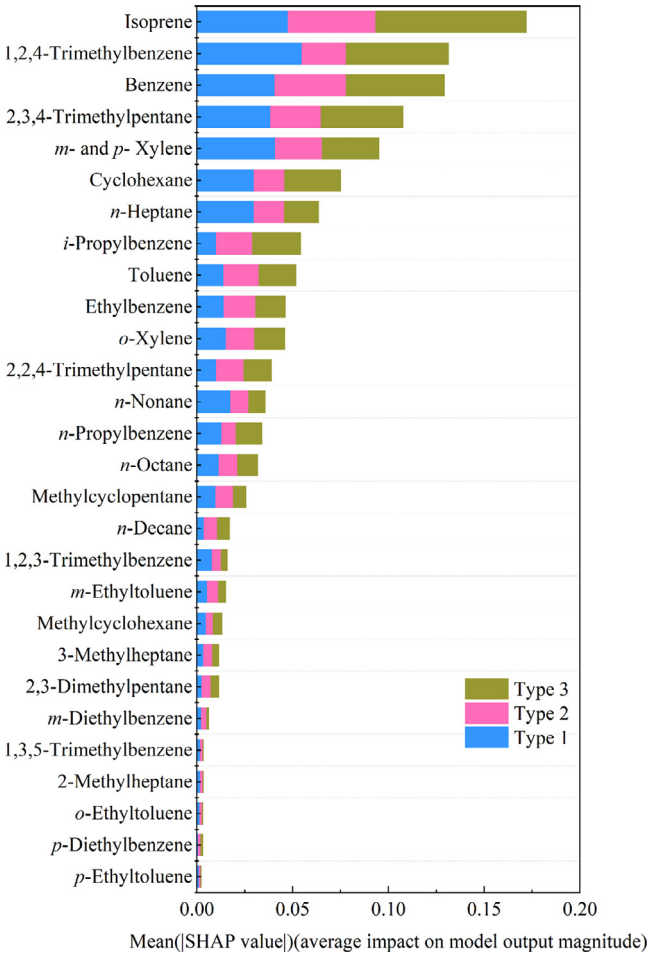

Fig. 4 – Stacked bar chart of average absolute Shapley additive explanation regression (SHAP) values of different variables. Type 1 represents the higher concentration of SOA with lower concentration of O_3; Type 2 represents the moderate concentration of SOA and O_3; Type 3 represents the higher concentration of O_3 with lower concentration of SOA.

mation studies have been widely conducted both on laboratory experiments and field measurements. In the ambient environment, SOA is mainly formed via gas-particle partitioning of SVOCs which is oxidation products of VOCs catalysed by hydroxyl radicals (•OH) Eq. (7)) (Srivastava et al., 2021; Wang et al., 2021). High temperature can enhance biogenic VOCs emission and promote the evaporation of semi-volatile SOA compounds to facilitate and inhibit SOA formation, respectively (Wu et al., 2020). In terms of O_3 formation, the OH-initiated degradation of VOCs results in RO_X ($RO_X = \cdot OH + \cdot HO_2 + \cdot RO_2$) radical cycle (Eqs. (8)–(11)) (Vermeuel et al., 2019; Wang et al., 2017), which led to complex relationship between O_3 and VOCs.

$$VOC_i + \cdot OH \xrightarrow{k_{OH,i}} SOA_i \quad (7)$$

$$VOCs + OH + O_2 \rightarrow RO_2 + H_2O \quad (8)$$

$$RO_2 + NO \rightarrow RO + NO_2 \quad (9)$$

$$RO + O_2 \rightarrow HO_2 + \text{carbonyls} \quad (10)$$

$$HO_2 + NO \rightarrow OH + NO_2 \quad (11)$$

The aforementioned complex competition mechanism illuminates that VOCs have strong photochemical reaction activity makes that photochemical consumptions of VOCs plays an important part in O_3 and SOA formation (Gao et al., 2019; Gligorovski et al., 2018; Li et al., 2020a). Observed VOCs concentration may be the residues of photochemical consumptions. To clarify this issue, we conducted on theoretical calculations for O_3 and SOA, and will elaborate on this in the next section.

2.3. *Reactive VOCs species to O_3 and SOA formation potential*

The calculated importances of $VOC_{consumed}$ to OFP and SOAFP during different seasons are shown in Fig. 6. The importance of consumed alkanes and alkyne both to OFP and SOAFP were very low during the observation in daytime owing to its low MIR and FAC value (Table S1) (Carter, 2010; Song et al., 2018). Consumed alkenes were the most dominant contributor to OFP, accounting for about 88%, especially in summer accounting for 94.03%. Among them shown in Fig. 7a, isoprene contributed the most to OFP at 41.45%, followed by *cis*-2-butene (16.70%), ethylene (11.03%), and propylene (10.27%), regardless of the season. Moreover, consumed aromatics have a minor but not negligible importance to OFP, accounting for 15.66%, 4.55%, 7.46% and 10.45% during different seasons, respectively. As for SOAFP, the importance of consumed aromatics and alkenes to SOAFP was comparable, accounting for 40.84% and 56.65% of the total SOAFP, as shown in Fig. 6b. The most recent study also found that the compositions of consumed VOCs were quite different that aromatics and alkenes dominated the consumed VOCs, accounting for 87% (Wang et al., 2020a). In winter, the most abundant SOAFP species were consumed aromatics, which contributed 58.27% of the total SOAFP. Especially in hazy days, the importance of aromatics to SOAFP is more than 90% of the measured VOC species in winter (Sun et al., 2016). Fig. 7b indicates that isoprene was also the dominant contributor over the year, accounting for 74.95% of the total SOAFP. While, *m*- and *p*-xylene, *o*-xylene, toluene, and ethylbenzene were determined to be the top 5 VOC SOAFP contributors, with respective proportions accounting for 10.59%, 5.20%, 4.86%, 1.20%, respectively. Consequently, consumed alkenes and aromatics have greater impact on OFP and SOAFP, respectively. These theoretical calculations suggest that alkenes are the dominant contributor and has great potential in the formation of ozone and the importance of aromatics to SOAFP is very important, which have verified the machine learning results. Therefore, we concluded that the observed VOCs are the residual VOCs after the severe photochemical reactions.

Fig. 8 illustrated the distribution of the top 7 VOCs that contribute to both O_3 and SOA. As is the case in machine learning results (Fig. 4), theoretical analysis results shows that isoprene, *m*- and *p*-xylene, 1,2,4-trimethylbenzene, toluene, and ethylbenzene are not only responsible for photochemical O_3 formation, but also important for SOA production. Therefore, the competition between photochemical reaction and

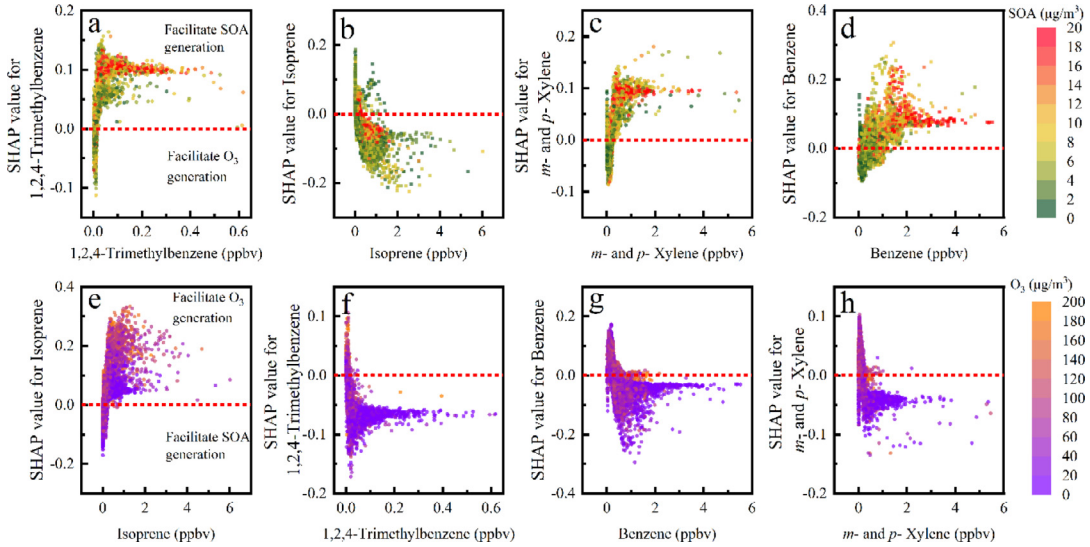

Fig. 5 – SHAP dependence plot of the top four most important variables for Type 1 and Type 3. The red dotted horizontal line indicates zero SHAP values. Ranked by SHAP values, in Type 1, from (a) to (d) are 1,2,4-trimethylbenzene, isoprene, *m*- and *p*-xylene and benzene; in Type 3, from (e) to (h) are isoprene, 1,2,4-trimethylbenzene, benzene and *m*- and *p*-xylene.

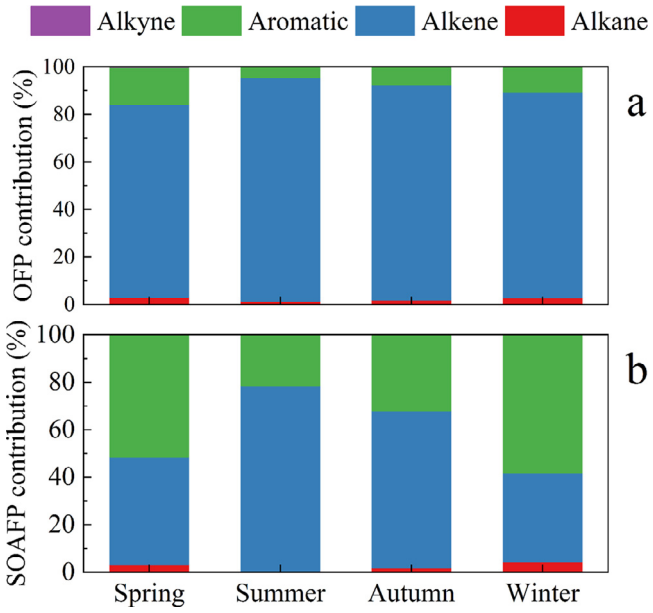

Fig. 6 – (a) Ozone formation potential (OFP) and (b) secondary organic aerosol formation potential (SOAFP) distributions (%) of VOC classes in spring, summer, autumn and winter.

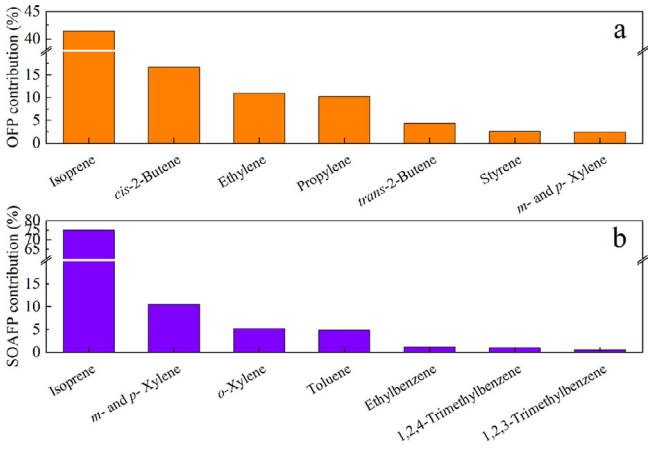

Fig. 7 – Top 7 greatest contribution to the total (a) OFP and (b) SOAFP.

aerosol chemical reactions of isoprene, *m*- and *p*-xylene, 1,2,4-trimethylbenzene, toluene, and ethylbenzene against OH radical impacts the formation O_3 and SOA.

3. Conclusions

A two-years observation of VOCs was obtained by continuously measurement from June 2017 to May 2019. VOCs and SOA concentrations are showing a certain regularity variation, being lower in summer and higher in winter; while O_3 concentration was completely opposite to those displayed by VOCs and SOA. Alkanes are the most abundant VOCs group, especially in summer, accounting for 66.79% of the total VOCs. Alkenes contribute the smallest proportion (9.11%) in summer while O_3 concentration (58.33 ppbv) is highest. Aromatics are abundant in summer contributing about 13.94%, with low SOA concentration. Disparities in VOCs proportions are observed responding to complex non-linear relationship of O_3 and SOA towards VOCs.

Machine learning method combining with theoretical calculations was found to differ substantially from the observation results, and revealed that observed VOCs are the residual VOCs after the severe photochemical reactions. The data-driven analysis results illustrated that alkenes (mainly ethy-

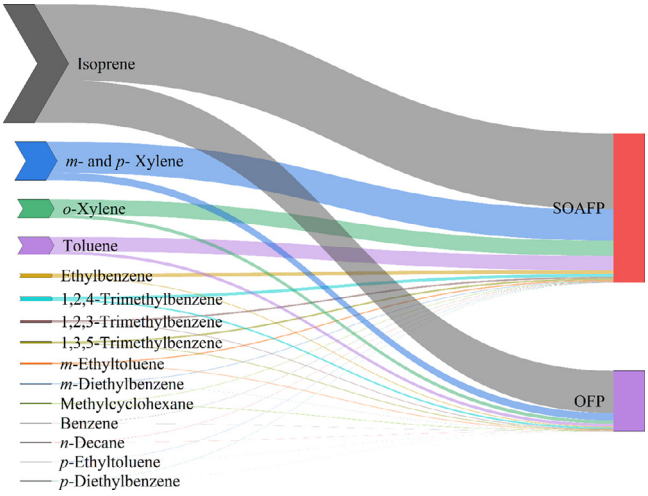

Fig. 8 – Schematic diagram for the distribution of the top ten VOCs that contribute to both O_3 and SOA.

lene, propylene, and isoprene) has the highest importance to O_3 formation (38%); alkanes (C_n, $n \geq 6$) and aromatics are the main source of SOA formation. The photochemical consumptions of VOCs were estimated. More than 80% of the VOCs consumed were alkenes which dominated the O_3 formation. Especially, isoprene contributed the most to OFP at 41.45%, regardless of the season. In terms of SOA formation, the importance of consumed aromatics (40.84%) and alkenes (56.65%) to SOAFP was comparable. Especially in winter, aromatics (58.27%) contributed the most to SOAFP. Collectively, alkenes are greatly depleted in summer with high OH reactivity relative stimulating ozone production, and exhibits low proportion. Aromatics played a leading role in SOA formation in winter.

In our study, machine learning method has the higher model accuracy (cross validation values) and is suitable for complex non-linear photochemistry and aerosol chemistry reaction of O_3 and SOA formation. In addition, to make up for the shortcomings of data-driven method in terms of uncertainty and interpretability, theoretical analysis should be suggested as a constraint in the application of machine learning method. Therefore, our findings provide scientific evidences to support the need for coordinated VOCs controls at seasonal and regional scales to achieve effective pollution reduction in both SOA and O_3. Moreover, the competition mechanisms responsible for O_3 and SOA formation against VOCs remain uncertain and is urgently needed.

Acknowledgments

This work was financially supported by the National Key Research and Development Program of China (No. 2018 YFE0106900). This study was supported by National Natural Science Foundation of China (Nos. 42077191, 41775149), Fundamental Research Funds for the Central Universities (No. 63213072), National Research Program for Key Issues in Air Pollution Control (No. DQGG-05-30) and the Blue Sky Foundation. This work is an importance from State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control.

Appendix A Supplementary data

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jes.2021.07.026.

REFERENCES

An, Z.S., Huang, R.J., Zhang, R.Y., Tie, X.X., Li, G.H., Cao, J.J., et al., 2019. Severe haze in northern China: a synergy of anthropogenic emissions and atmospheric processes. Proc. Natl. Acad. Sci. USA 116, 8657–8666.

Buysse, C.E., Kaulfus, A., Nair, U., Jaffe, D.A., 2019. Relationships between particulate matter, ozone, and nitrogen oxides during urban smoke events in the Western US. Environ. Sci. Technol. 53, 12519–12528.

Cardelino, C.A., Chameides, W.L., 1995. An observation-based model for analyzing ozone precursor relationships in the urban atmosphere. J. Air Waste Manag. Assoc. 45, 161–180.

Carter, W.P.L., 2010. Development of a condensed SAPRC-07 chemical mechanism. Atmos. Environ. 44, 5336–5345.

Chen, H.M., Zhuang, B.L., Liu, J., Wang, T.J., Li, S., Xie, M., et al., 2019. Characteristics of ozone and particles in the near-surface atmosphere in the urban area of the Yangtze River Delta, China. Atmos. Chem. Phys. 19, 4153–4175.

Choubin, B., Abdolshahnejad, M., Moradi, E., Querol, X., Mosavi, A., Shamshirband, S., et al., 2020. Spatial hazard assessment of the PM_{10} using machine learning models in Barcelona, Spain. Sci. Total Environ. 701, 134474.

Colmer, J., Hardman, I., Shimshack, J., Voorheis, J., 2020. Disparities in $PM_{2.5}$ air pollution in the United States. Science 369, 575–578.

Dai, Q.L., Liu, B.S., Bi, X.H., Wu, J.H., Liang, D.N., Zhang, Y.F., et al., 2020. Dispersion normalized PMF provides insights into the significant changes in source contributions to $PM_{2.5}$ after the COVID-19 Outbreak. Environ. Sci. Technol. 54, 9917–9927.

Gao, J., Wei, Y.T., Shi, G.L., Yu, H.F., Zhang, Z.C., Song, S.J., et al., 2020. Roles of RH, aerosol pH and sources in concentrations of secondary inorganic aerosols, during different pollution periods. Atmos. Environ. 241, 117770.

Gao, Y.Q., Wang, H.L., Zhang, X., Jing, S.A., Peng, Y.R., Qiao, L.P., et al., 2019. Estimating secondary organic aerosol production from toluene photochemistry in a megacity of China. Environ. Sci. Technol. 53, 8664–8671.

Gligorovski, S., Strekowski, R., Barbati, S., Vione, D., 2018. Addition and correction to environmental implications of hydroxyl radicals (•OH). Chem. Rev. 118 2296-2296.

Grosjean, D., 1992. In situ organic aerosol formation during a smog episode: Estimated production and chemical functionality. Atmos. Environ. 26, 953–963.

Grosjean, D., Seinfeld, J.H., 1989. Parameterization of the formation potential of secondary organic aerosols. Atmos. Environ. 23, 1733–1747.

Gu, Y., Liu, B.S., Li, Y.F., Zhang, Y.F., Bi, X.H., Wu, J.H., et al., 2020. Multi-scale volatile organic compound (VOC) source apportionment in Tianjin, China, using a receptor model coupled with 1-hr resolution data. Environ. Pollut. 265, 115023.

Gu, Y.Y., Li, Q.Q., Wei, D., Gao, L.R., Tan, L., Su, G.J., et al., 2019. Emission characteristics of 99 NMVOCs in different seasonal days and the relationship with air quality parameters in Beijing, China. Ecotoxicol. Environ. Saf. 169, 797–806.

Huang, S., Shao, M., Lu, S.H., Liu, Y., 2008. Reactivity of ambient volatile organic compounds (VOCs) in summer of 2004 in Beijing. Chin. Chem. Lett. 19, 573–576.
Kang, M.J., Zhang, J., Zhang, H.L., Ying, Q., 2021. On the relevancy of observed ozone increase during COVID-19 Lockdown to summertime ozone and $PM_{2.5}$ control policies in China. Environ. Sci. Technol. Lett. 8, 289–294.
Li, J., Han, Z.W., Sun, Y.L., Li, J.W., Liang, L., 2021. Chemical formation pathways of secondary organic aerosols in the Beijing-Tianjin-Hebei region in wintertime. Atmos. Environ. 244, 117996.
Li, K., Jacob, D.J., Liao, H., Zhu, J., Shah, V., Shen, L., et al., 2019. A two-pollutant strategy for improving ozone and particulate air quality in China. Nat. Geosci. 12, 906–910.
Li, N., Jiang, Q., Wang, F.S., Xie, J., Li, Y.Y., Li, J.S., et al., 2020a. Emission behavior, environmental impact and priority-controlled pollutants assessment of volatile organic compounds (VOCs) during asphalt pavement construction based on laboratory experiment. J. Hazard. Mater. 398, 122904.
Li, Q.Q., Su, G.J., Li, C.Q., Liu, P.P., Zhao, X.X., Zhang, C.L., et al., 2020b. An investigation into the role of VOCs in SOA and ozone production in Beijing, China. Sci. Total Environ. 720, 137536.
Li, Z.Y., Yim, S.H.-L., Ho, K.-F., 2020c. High temporal resolution prediction of street-level $PM_{2.5}$ and NOx concentrations using machine learning approach. J. Clean. Prod. 268, 121975.
Liu, Y., Shao, M., Fu, L.L., Lu, S.H., Zeng, L.M., Tang, D.G., 2008. Source profiles of volatile organic compounds (VOCs) measured in China: part I. Atmos. Environ. 42, 6247–6260.
Liu, Y.F., Kong, L.W., Liu, X.G., Zhang, Y.P., Li, C.L., Zhang, Y.Y., et al., 2021. Characteristics, secondary transformation, and health risk assessment of ambient volatile organic compounds (VOCs) in urban Beijing, China. Atmos. Pollut. Res. 12, 33–46.
Lu, X., Hong, J.Y., Zhang, L., Cooper, O.R., Schultz, M.G., Xu, X.B., et al., 2018. Severe surface ozone pollution in China: a global perspective. Environ. Sci. Technol. Lett. 5, 487–494.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS). Long Beach, CA, USA.
Meng, Z., Dabdub, D., Seinfeld, J.H., 1997. Chemical coupling between atmospheric ozone and particulate matter. Science 277, 116–119.
Min, S., Bin, W., Sihua, L., Bin, Y., Ming, W., 2011. Effects of Beijing olympics control measures on reducing reactive hydrocarbon species. Environ. Sci. Technol. 45, 514–519.
Niu, H., Mo, Z.W., Shao, M., Lu, S.H., Xie, S.D., 2016. Screening the emission sources of volatile organic compounds (VOCs) in China by multi-effects evaluation. Front. Environ. Sci. Eng. 10, 1.
Peng, J.F., Hu, M., Shang, D.J., Wu, Z.J., Du, Z.F., Tan, T.Y., et al., 2021. Explosive secondary aerosol formation during severe haze in the North China Plain. Environ. Sci. Technol. 55, 2189–2207.
Qin, J.J., Wang, X.B., Yang, Y.R., Qin, Y.Y., Shi, S.X., Xu, P.H., et al., 2021. Source apportionment of VOCs in a typical medium-sized city in North China Plain and implications on control policy. J. Environ. Sci. 107, 26–37.
Shang, D.J., Peng, J.F., Guo, S., Wu, Z.J., Hu, M., 2020. Secondary aerosol formation in winter haze over the Beijing-Tianjin-Hebei Region, China. Front. Environ. Sci. Eng. 15, 34.
Song, M.D., Tan, Q.W., Feng, M., Qu, Y., Liu, X.G., An, J.L., et al., 2018. Source apportionment and secondary transformation of atmospheric nonmethane hydrocarbons in Chengdu, Southwest China. J. Geophys. Res. Atmos. 123, 9741–9763.
Srivastava, D., Daellenbach, K.R., Zhang, Y., Bonnaire, N., Chazeau, B., Perraudin, E., et al., 2021. Comparison of five methodologies to apportion organic aerosol sources during a PM pollution event. Sci. Total Environ. 757, 143168.
Sun, J., Wu, F.K., Hu, B., Tang, G.Q., Zhang, J.K., Wang, Y.S., 2016. VOC characteristics, emissions and contributions to SOA formation during hazy episodes. Atmos. Environ. 141, 560–570.
Tan, Z.F., Lu, K.D., Dong, H.B., Hu, M., Li, X., Liu, Y.H., et al., 2018. Explicit diagnosis of the local ozone production rate and the ozone-NOx-VOC sensitivities. Sci. Bull. 63, 1067.
Tian, Y.Z., Zhang, Y.F., Liang, Y.L., Niu, Z.B., Xue, Q.Q., Feng, Y.C., 2020. $PM_{2.5}$ source apportionment during severe haze episodes in a Chinese megacity based on a 5-month period by using hourly species measurements: Explore how to better conduct PMF during haze episodes. Atmos. Environ. 224, 117364.
Vermeuel, M.P., Novak, G.A., Alwe, H.D., Hughes, D.D., Kaleel, R., Dickens, A.F., et al., 2019. Sensitivity of ozone production to NOx and VOC along the Lake Michigan coastline. J. Geophys. Res. Atmos. 124, 10989–11006.
Wang, H.L., Chen, C.H., Wang, Q., Huang, C., Su, L.Y., Huang, H.Y., et al., 2013a. Chemical loss of volatile organic compounds and its impact on the source analysis through a two-year continuous measurement. Atmos. Environ. 80, 488–498.
Wang, H.L., Wang, Q., Gao, Y.Q., Zhou, M., Jing, S.G., Qiao, L.P., et al., 2020a. Estimation of secondary organic aerosol formation during a photochemical smog episode in Shanghai, China. J. Geophys. Res. Atmos. 125, e2019JD032033.
Wang, J.F., Ye, J.H., Zhang, Q., Zhao, J., Wu, Y.Z., Li, J.Y., et al., 2021. Aqueous production of secondary organic aerosol from fossil-fuel emissions in winter Beijing haze. Proc. Natl. Acad. Sci. USA 118, e2022179118.
Wang, M., Shao, M., Lu, S.-H., Yang, Y.-D., Chen, W.-T., 2013b. Evidence of coal combustion contribution to ambient VOCs during winter in Beijing. Chin. Chem. Lett. 24, 829–832.
Wang, T., Xue, L.K., Brimblecombe, P., Lam, Y.F., Li, L., Zhang, L., 2017. Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects. Sci. Total Environ. 575, 1582–1596.
Wang, Y.H., Gao, W.K., Wang, S., Song, T., Gong, Z.Y., Ji, D.S., et al., 2020b. Contrasting trends of $PM_{2.5}$ and surface-ozone concentrations in China from 2013 to 2017. Natl. Sci. Rev. 7, 1331–1339.
Wu, K., Yang, X.Y., Chen, D., Gu, S., Lu, Y.Q., Jiang, Q., et al., 2020. Estimation of biogenic VOC emissions and their corresponding impact on ozone and secondary organic aerosol formation in China. Atmos. Res. 231, 104656.
Xie, Y.R., Castro, D.C., Bell, S.E., Rubakhin, S.S., Sweedler, J.V., 2020. Single-cell classification using mass spectrometry through interpretable machine learning. Anal. Chem. 92, 9338–9347.
Yang, Y.C., Liu, X.G., Zheng, J., Tan, Q.W., Feng, M., Qu, Y., et al., 2019. Characteristics of one-year observation of VOCs, NOx, and O_3 at an urban site in Wuhan, China. J. Environ. Sci. 31, 297–310.
Zhang, D., He, B., Yuan, M.H., Yu, S.J., Yin, S.S., Zhang, R.Q., 2021. Characteristics, sources and health risks assessment of VOCs in Zhengzhou, China during haze pollution season. J. Environ. Sci. 108, 44–57.
Zhang, K., Xu, J.L., Huang, Q., Zhou, L., Fu, Q.Y., Duan, Y.S., et al., 2020a. Precursors and potential sources of ground-level ozone in suburban Shanghai. Front. Environ. Sci. Eng. 14, 92.
Zhang, Q., Zheng, Y.X., Tong, D., Shao, M., Wang, S.X., Zhang, Y.H., et al., 2019a. Drivers of improved $PM_{2.5}$ air quality in China from 2013 to 2017. Proc. Natl. Acad. Sci. USA 116, 24463–24469.
Zhang, X.F., Gao, S., Fu, Q.Y., Han, D.M., Chen, X.J., Fu, S., et al., 2020b. Impact of VOCs emission from iron and steel industry

on regional O_3 and $PM_{2.5}$ pollutions. Environ. Sci. Pollut. Res. 27, 28853–28866.
Zhang, Y.M., Vu, T.V., Sun, J.Y., He, J.J., Shen, X.J., Lin, W.L., et al., 2019b. 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.
Zheng, H., Kong, S.F., Yan, Q., Wu, F.Q., Cheng, Y., Zheng, S.R., et al., 2019. The impacts of pollution control measures on $PM_{2.5}$ reduction: Insights of chemical composition, source variation and health risk. Atmos. Environ. 197, 103–117.
Zou, Y., Deng, X.J., Zhu, D., Gong, D.C., Wang, H., Li, F., et al., 2015. Characteristics of 1 year of observational data of VOCs, NOx and O_3 at a suburban site in Guangzhou, China. Atmos. Chem. Phys. 15, 6625–6636.