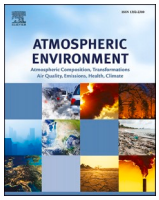
Atmospheric Environment 334 (2024) 120714

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv



Combined PMF modelling and machine learning to identify sources and meteorological influencers of volatile organic compound pollution in an industrial city in eastern China

Wei Chen [a,b], Xuezhe Xu [a,*], Wenqing Liu [a]

[a] Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, 230031, China
[b] Science Island Branch, Graduate School, University of Science and Technology of China, Hefei, 230026, China

HIGHLIGHTS

- Explored the impacts of driving factors on VOC concentrations in Huaibei.
- Employed a machine learning method (CatBoost-SHAP) combined with the PMF model.
- TVOCs concentrations varied from 6.0 to 64.9 ppb, lowest in summer and highest in winter.
- Emissions, mainly from industry (SS) and vehicles (VE), contributed 42.5% to TVOCs.

ARTICLE INFO

Keywords:
Volatile organic compounds
Machine learning
SHAP
Positive matrix factorization
Meteorology

ABSTRACT

The concentrations of volatile organic compounds (VOCs) in the atmosphere are driven by both emissions and meteorology. Additionally, a non-linear relationship has been sown between VOCs and the factors driving their concentrations, warranting further investigation into this relationship. Based on a combination of machine learning and Shapley Additive exPlanation (SHAP) models, the aim of this study is to characterize relationships among VOC concentration, meteorology, atmospheric oxidation, and positive matrix factorization (PMF) source apportionment data for Huaibei in 2022. The concentration of total VOCs (TVOCs) in 2022 was observed to be 6.0–64.9 ppb, with the lowest and highest concentrations being observed in the summer and winter, respectively. The receptor model identified vehicle exhaust (VE) as being the highest contributor to TVOCs, whereas biogenic sources (BS) were marked as being the lowest. Moreover, the effect of the emissions on TVOCs was 42.5%, with solvent sources (SS) and VE being the main sources of VOCs, accounting for approximately 45.8%, whereas meteorological conditions and atmospheric oxidation capacity contributed 50.9% and 6.6%, respectively. Temperature (T) and relative humidity (RH) accounted for more than 60% of the meteorological effects, though this trend varied across seasons. RH in spring and autumn had the greatest impact on VOC concentrations, whereas T was the main factor in summer. Finally, the contribution of VE was the highest among the emission sources in the summer, whereas that of SS was the highest in the winter. Overall, this study showed the feasibility of combining machine learning and SHAP models to elucidate the dynamics of VOCs based on observed data.

1. Introduction

As critical precursors of O_3 and $PM_{2.5}$ (Tan et al., 2021; Liu and Shi, 2021), volatile organic compounds (VOCs) are the main determinants of air quality. Therefore, continuous reduction of VOC emissions is of utmost importance in the context of improving China (Guan et al., 2023; Hang et al., 2023; Wang et al., 2024).

VOCs have been shown to stem from numerous sources, including vehicles exhaust (Drozd et al., 2016), coal and biomass combustion (Cheng et al., 2018; Wang et al., 2014), gasoline evaporation (Liu et al., 2015), the application of paints and organic solvents (Yuan et al., 2010), industrial processes (Mo et al., 2017), and biogenic emissions (Liang et al., 2020; Wang et al., 2020; Mozaffar and Zhang, 2020). In addition, the measured concentrations of VOCs in the atmosphere are influenced

* Corresponding author.
E-mail address: xzxu@aiofm.ac.cn (X. Xu).

https://doi.org/10.1016/j.atmosenv.2024.120714
Received 21 March 2024; Received in revised form 7 July 2024; Accepted 22 July 2024
Available online 23 July 2024

by meteorological factors such as temperature (T), relative humidity (RH), the intensity of radiation, and total oxidant (Ox, O_3, and NO_2) concentrations (Wang et al., 2022b). Meteorology affects the emissions, accumulation, and diffusion of VOCs (Hunter-Sellars et al., 2020; Li et al., 2020; Markowicz and Larsson, 2015; Nussbaumer and Cohen, 2020; Wang et al., 2022a), and it determines the rates at which the photochemical reactions of VOCs occur (Hu et al., 2021; Liu et al., 2022b; Peng et al., 2021). Thus, it is necessary to identify and evaluate the contribution of emission and meteorological factors to atmospheric VOCs, since these steps could yield novel insights that could be applied in the context of significantly improving air quality.

The PMF model has been extensively used to identify sources of VOC emissions (Dai et al., 2021; Fu et al., 2020b; Gao et al., 2020; Guan et al., 2023; Qin et al., 2021). However, this model has not been the most effective solution in the context of quantifying the effects of source emissions and meteorology on VOC concentration. In addition, many studies have been conducted on the relationships between VOCs and various driving factors; to this end, chemical transport models (CTMs) and statistical methods are commonly used tools (Guo et al., 2022; Li et al., 2019; Nelson et al., 2023; Zhang et al., 2023). However, these models have several limitations. For example, CTMs require detailed input data and are computationally intensive (Nelson et al., 2023); additionally, due to their reliance on simplified assumptions and parameters, these models tend to not accurately and reliably characterize the dynamics of air pollutants (Sayeed et al., 2022). Moreover, these traditional models may oversimplify the complex relationships between air pollutant concentrations and certain variables (Peng et al., 2023; Reid et al., 2015). In contrast, machine learning (ML) approaches have been highlighted as being highly effective in the resolution of nonlinear problems(Nelson et al., 2023). However, despite the demonstrated effectiveness of ML models, they tend to have a “black box” nature, making the interpretation and generalizability (i.e., to real scenarios) of their results difficult (Liu et al., 2022a). The SHapley Additive explanation (SHAP) algorithm has been applied to resolve these concerns (Lundberg and Lee, 2017; Sadeghi et al., 2022; Nelson et al., 2023).

To date, studies using ML-coupled receptor models have been applied to identify and quantify the impacts of driving factors on pollutants, including $PM_{2.5}$ and O_3 (Hou et al., 2022; Wang et al., 2022a; Xu et al., 2023; Zhang et al., 2022). For instance, Hou et al. (2022) employed a Random Forest (RF) model coupled with a Shapley additive explanation algorithm to investigate the roles of major meteorological factors, primary emissions, and chemistry in the dynamics of $PM_{2.5}$ pollution; Zhang et al. (2022) revealed the effects of sources and meteorological factors on $PM_{2.5}$ pollution using the RF method combined with a positive matrix factorization (PMF) receptor model; and Xu et al. (2023) used ML models to analyze the effects of emissions and meteorological factors on CO_2 and $PM_{2.5}$ generation. In contrast, Nelson et al. (2023) have highlighted the contributions of local emissions and meteorological conditions to O_3 by integrating the PMF with RF-SHAP models. Nevertheless, in the ambient air of China, there is a scarcity of studies that have used ML-coupled receptor models to evaluate the impact of various factors (e.g., emissions and meteorological conditions) on VOCs.

Therefore, the aim of this study is to quantify the effects of driver factors on VOCs in the industrial city of Huaibei, Anhui Province, in eastern China, in 2022. These effects will be characterized based on observational data on pollutant concentrations, meteorological factors, and VOC composition. This study was conducted as follows: we first used the PMF model to analyze the sources of VOCs; we used the ML model to study the effects of the main drivers; and finally, the quantitative impact of each driving factor on the predicted TVOCs concentrations was determined via the SHAP model. This study provides a framework for the exploration of the effects of co-driving factors in VOCs, laying a scientific foundation for the continued improvement of air quality in China.

2. Materials and methods

The research methodology framework used in this study is shown in Fig. 1. First, we conducted the source apportionment of VOCs based on the methods recommended by Wang et al. (2022a) and Xu et al. (2023). Second, the ML model was used to study the effects of the main drivers (i.e., source apportionment results, meteorological factors, and atmospheric oxidation) (Zhang et al., 2022) on TVOCs levels. The SHAP values for each driving factor were obtained from the SHAP model (Hou et al., 2022). Subsequently, the quantitative impact of each driving factor on the predicted TVOCs concentrations (SHAP baseline) was determined. The detailed methods are as follows.

2.1. Study area and data details

Huaibei (116° 24′–117° 03′ E, 33° 16′–34° 10′ N), an important, resource-based city with a total area of 2741 km^2, is located in the north of Anhui Province, China, interfacing Jiangsu and Henan provinces (Fig. 2). Administrative base map data of this region were obtained from the China Basic Geographic Information Database, and elevation data were obtained from the Geospatial Data Cloud (https://www.gscloud.cn/); GDEMV2 30 m resolution data were released.

The pollutant concentration data used in this study were obtained hourly from the National Urban Environmental Air Automatic Monitoring System of the China Environmental Monitoring Station (https://air.cnemc.cn:18007/). The data included hourly concentrations of O_3 and NO_2 in Huaibei in 2022, which were used to calculate the total oxidant concentration ($Ox = O_3+NO_2$) (Fu et al., 2020a; Zhao et al., 2020). The meteorological element data was acquired from the Huaibei Meteorological Station (116° 50′ E, 33° 59′ N, 31.5 m), whereas the VOCs composition observation data was acquired from the Huaibei Photochemical Composition Station (116° 48′ 1″ E, 33° 58′ 30″ N, altitude 44.6 m); this station is located on the roof of the Huaibei Monitoring Station. No obvious industrial emission sources were found in the surrounding areas, ensuring objective reflection of the air quality status of Huaibei. Surface net solar radiation (SSR, W/m^2) data were obtained from reanalysis data of the ERA5 dataset (https://cds.climate.copernicus.eu/cdsapp#/Dataset/reanalysis era, five single levels? Tab = form). Meteorological parameters include surface pressure (SP, kPa), RH (%), T (°C), WD (°), and WS (m/s).

The concentrations of VOCs were monitored using an online gas chromatography system equipped with a mass spectrometer and a flame ionization detector (GC-MS/FID) (Super lab 2020-TT-GCMS, Shanghai Panhe Scientific Instrument Co., Ltd., China). Throughout the data acquisition process, the GC-MS/FID was calibrated daily to ensure consistency and sensitivity using internal and external standard gases. The VOC Standards from the U.S. EPA PAMS, TO-15, and 13 types of aldehyde and ketone standard mixtures were used for calibration. During the observation period, zero and span gas checks (PAMS calibration gases) were conducted monthly using the six-point method, together with the adjustment of the retention time. Additionally, the concentration-response standard curve contained six concentration levels (0.5×10^{-9}, 2×10^{-9}, 4×10^{-9}, 6×10^{-9}, 8×10^{-9}, and 10×10^{-9}) with the determinability coefficient (R^2) above 0.98. The specific species monitored, the concentration-response standard curve, and the statistics linked to the accuracy of the acquired data are provided in Tables S1 and S2. Bromochloromethane, 1,4-difluorobenzene, chlorobenzene, and 4-bromofluorobenzene were used as internal standards for MS calibration. In addition, external standard methods were used for the quantification of C2–C5 compounds, and internal standard methods were used for the quantification of C5–C12 compounds. Throughout the study period, 8683 sets of VOC data were obtained, including 106 VOC species (29 alkanes, 11 olefins, 1 alkyne, 17 aromatics, 35 halohydrocarbons, 12 oxygen-containing compounds (OVOCs), and one sulfur compound). The method detection limits (MDLs) of the instrument for all species ranged from 0.021 to 0.142 ppbv (Table S3).

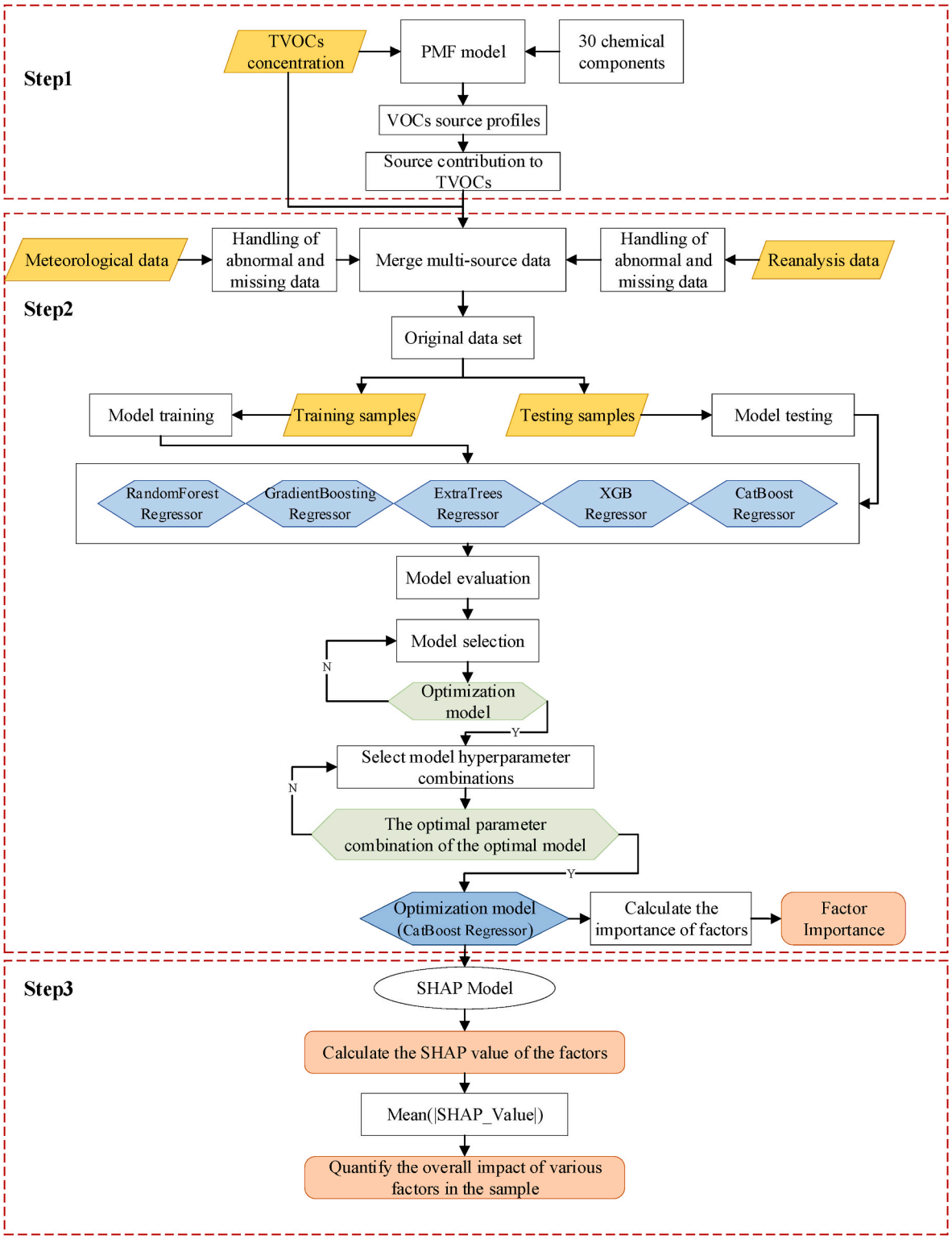


Fig. 1. Technology road map.

2.2. *Positive matrix factorization (PMF) model*

A statistical method, namely, PMF, was used to obtain the source contributions (as input data in the ML model) to VOCs. In this study, the total measured VOC was included as an input variable in the PMF model to directly obtain the source contribution. In addition, we adopted the following principles to guide our selection of the VOC species: (1) species with more than 25 % of data missing or below the MDLs were rejected, which follows the methodology of previous studies (Ling and Guo, 2014; Yu et al., 2022); (2) species with short atmospheric lifetimes were excluded due to their rapid diffusion upon emission into the atmosphere; (3) species that represent source tracers of emission sources were retained (e.g., in the case of isoprene); (4) the missing data were replaced by −999 or median values, whereas values below MDLs were used to increase the uncertainty. Eventually, 30 VOC species were selected for source apportionment analysis. VOC species were grouped

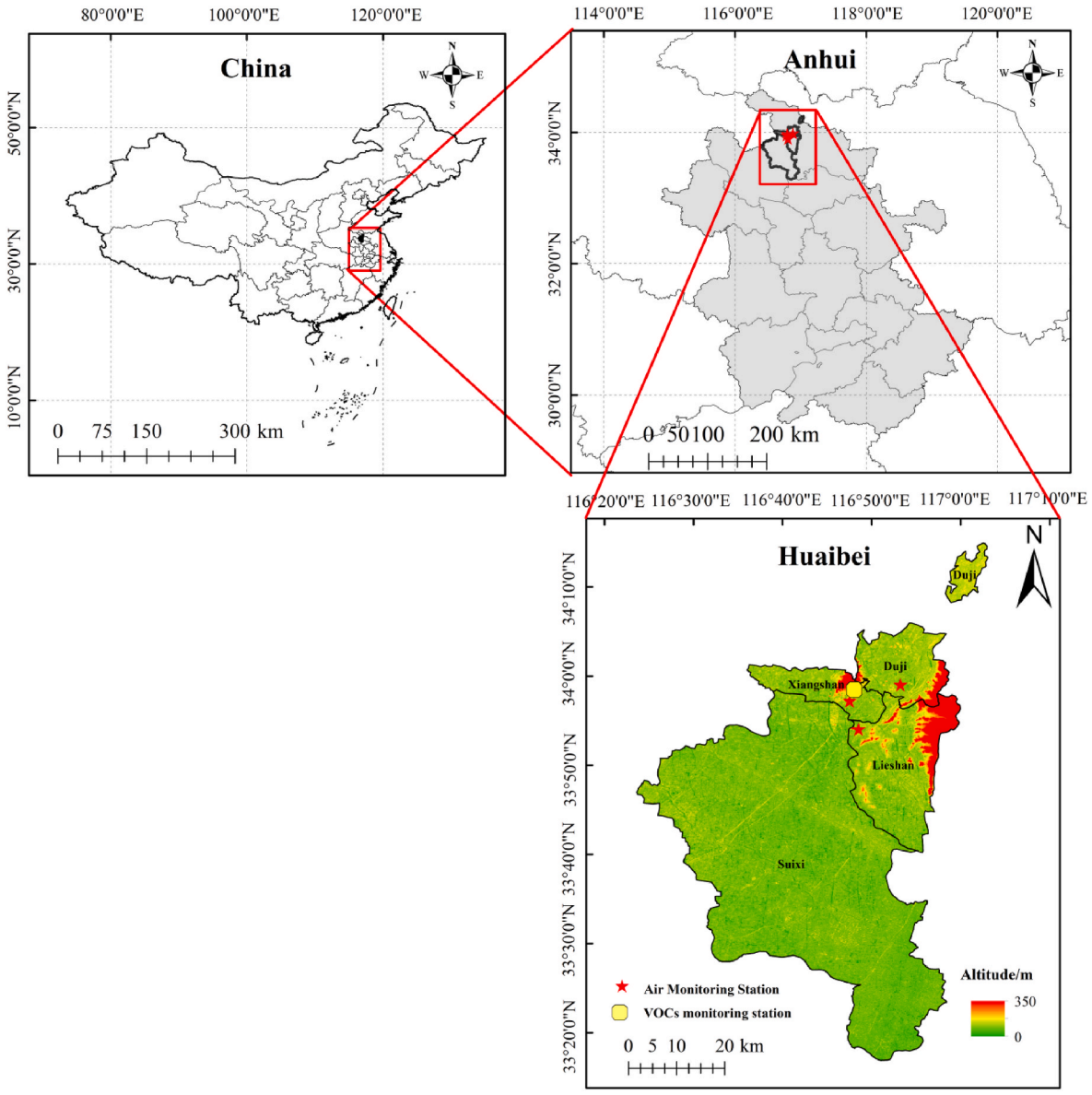


Fig. 2. The topography and distribution of study sites across the study area.

into strong, weak, and bad according to their signal/noise ratio (S/N), and there were 25 and five species grouped into strong and weak, respectively.

PMF is an analysis tool based on multivariate factors that decompose a matrix of specific sample data into two matrices, namely, factor contributions (G) and factor profiles (F) (Wang et al., 2022b; Wu et al., 2023a; Zheng et al., 2021). The principle underlying the PMF model can be described as follows (Paatero, 1997; Paatero and Tapper, 1994):

$$X_{ij} = \sum_{k=1}^{p} G_{ik} \times F_{kj} + E_{ij} \quad \text{(Eq. 1)}$$

where X_{ij} is the concentration of jth species in ith sample, G_{ik} is the contribution of the kth source to ith sample, F_{kj} is the jth species fraction from kth sources, E_{ij} is the residue factor of the jth species in ith sample, and p is the number of sources (Ling et al., 2011; Paatero, 1997). PMF minimizes the objective function Q:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{E_{ij}}{\mu_{ij}} \right)^2 \quad \text{(Eq. 2)}$$

where n and m are the number of species and samples, and μ_{ij} is the uncertainty of the jth species in the ith sample.

Uncertainty is calculated as follows (Eq. (3)):

$$\mu_{ij} = \begin{cases} \frac{5}{6} \times MDL, & \mu_{ij} \le \text{MDL} \\ \sqrt{(\text{RSD} \times X_{ij})^2 + MDL^2}, & \mu_{ij} > \text{MDL}, \end{cases} \quad \text{(Eq. 3)}$$

where RSD is the relative standard deviation of components, and MDL is the method detection limit (details are provided in Table S3). In this study, the program was run multiple times to find the smallest value of Q and observed the value of the residual error matrix E; optimizing for the smallness of this value ensured the accuracy of the simulation outcomes and their correlation with observations.

The results of bootstrap (BS), displacement (DISP), and BS-DISP were subsequently analyzed.

2.3. Machine learning (ML) model selection

To explore the data features, ML model training and evaluation were conducted using data from May 2021 to June 2023 in Huaibei. In this

study, the hourly data of T, RH, P, WS, WD, SSR, Ox, and TVOCs concentrations and the results of PMF were inputted into the ML model. Then the ML model, in conjunction with the SHAP model, was applied to obtain the SHAP values of driving factors as a quantitative impact of driving factors on the TVOCs prediction mean value. R2 was used as an evaluation indicator of model performance. The model with the highest R^2 and its corresponding parameter combination was selected from five ML models as the optimal ML model. The five tree-based ML models were as follows: RandomForestRegressor, GradientBoostingRegressor, ExtraTreesRegressor, XGB Regressor, and CatBoostRegressor. The RandomForestRegressor, GradientBoostingRegressor, and ExtraTreesRegressor were imported from the ensemble package called sklearn, whereas the XGB Regressor and CatBoostRegressor were imported from the XGBoost and CatBoost packages, respectively. The model parameters included the number of estimators and the depth of the tree. The calculation of R^2 is shown in Equation (3):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \quad \text{(Eq. 3)}$$

where n is the number of samples, y_i is the ith observational value, $\hat{y}_i$ is the ith predicted value, and $\overline{y}_i$ is the mean value of observational data.

In this study, datasets representative of the whole of 2022 and its various seasons, i.e., spring (March to May 2022), summer (June to August 2022), autumn (September to November 2022), and winter (January, February, and December 2022), were used for training and evaluation of the CatBoostRegressor. The data sets were randomly divided into training and testing sets in a ratio of 8:2, with corresponding input data structures of (5890,14) and (1472,14) for the full year of 2022, (1461,14) and (365,14) for the spring, (1431,14) and (358,14) for the summer, (1454,14) and (363,14) for the autumn, and (1544,14) and (386,14) for the winter. In the process of adjusting the hyperparameters of the machine learning model, several sets of quite different hyperparameters were selected for testing based on experience to determine the range of values. Then, experiments are conducted at points within this range to pick the best hyperparameter combination. The evaluation results for the five ML models under different parameter combinations are listed in Table 1. When *estimators* = 640 and *depth* = 10, the R^2 of the CatBoost regression was the highest (0.716). Therefore, CatBoost regression was combined with the SHAP model to evaluate the impact of driving factors on TVOCs.

CatBoost is a GBDT model based on symmetric decision trees that solve the difficult problem of categorical features by generating new numerical features based on the frequency of appearance and hyperparameters. CatBoost also solves the problems of model prediction and gradient bias, significantly improving the algorithm's generalization ability and accuracy (Huang et al., 2019). CatBoost performs excellently in ML tasks involving classification and heterogeneous data (Hancock and Khoshgoftaar, 2020; Jabeur et al., 2021), and has shown significant favorable accuracy, stability, and computational cost (Bo et al., 2022). CatBoost has been widely applied to reconstruct the annual average concentrations of CO_2 (Wu et al., 2023b), $PM_{2.5}$, and O_3 with a spatial resolution of 21 km × 1 km (Wang et al., 2022c, 2023). In developing technology for predicting atmospheric pollutants, the combination of vector autoregression (VAR), the Kriging method, and XGBoost (Extreme Gradient Boost) can predict the spatial concentration distribution of O_3 (Dai et al., 2023). In a previous study, LightGBM, Random Forest, Catboost, Adaboost, and XGBoost were applied to obtain the air quality index (AQI) of an Indian city in September 2022, and the results showed that CatBoost had the best performance (Ravindiran et al., 2023). Similarly, in this study, a CatBoost Regression model was constructed using the Python model library CatBoost.

2.4. Shapley Additive exPlanation (SHAP) model

To clarify the relationship between each variable and the complex prediction, the SHAP method, which distributes the total gains among players based on the Sharpley value, was applied (Lundberg and Lee, 2017; Sadeghi et al., 2022; Nelson et al., 2023). As the SHAP model was proposed by Lundbegr and Lee (2017), it has been widely applied in exploring the influence of driving factors in haze events (Hou et al., 2022), revealing the effects of emissions and meteorology on $PM_{2.5}$ (Zhang et al., 2022), and adjusting cloud fraction of aerosols and their dependence on meteorological control (Jia et al., 2023). In this study, the SHAP package of Python was applied, whereas TreeExplainer's model interpreter was used to calculate the impact of various features on the prediction results of the ML models.

Generally, the relationship between the SHAP and predicted values is defined by Equation (4).

$$f(x) = \phi_0(f,x) + \sum_{i=1}^{p}\phi_i(f,x) \quad (4)$$

where $f(x)$ represents the predicted value of the model; ϕ_0 represents the baseline value of the model, which is the mean of all predicted values; p represents the number of factors, $\phi_i(f,x)$ is the SHAP value of the ith driving factor, indicating the impact of the ith driving factor.

3. Results and discussion

3.1. Observational data

Surface pressure (SP), relative humidity (RH), temperature (T), solar surface radiation (SSR), and total oxidant concentration (Ox) for 2022 in Huaibei are shown in Fig. S1, and wind rose charts for different seasons are shown in Fig. S2. The annual average values of SP, RH, T, SSR, and Ox were 1012.81 kPa, 67.99%, 16.03 °C, 511.92 kW/m^2, and 97.76 μg/

Table 1
Comparison of five tree-based ML models based on predictive performance.

R^2	*estimators*	160					320				
	depth	8	9	10	11	12	8	9	10	11	12
RandomForestRegressor		0.522	0.552	0.576	0.594	0.576	0.522	0.553	0.574	0.598	0.578
GradientBoostingRegressor		0.675	0.682	0.694	0.688	0.655	0.685	0.687	0.698	0.691	0.655
ExtraTreesRegressor		0.466	0.497	0.529	0.556	0.557	0.465	0.502	0.53	0.558	0.558
XGBRegressor		0.662	0.639	0.659	0.642	0.637	0.664	0.641	0.66	0.642	0.637
CatBoostRegressor		0.657	0.68	0.681	0.679	0.655	0.692	0.693	0.702	0.707	0.687
R^2	*estimators*	480					640				
	depth	8	9	10	11	12	8	9	10	11	12
RandomForestRegressor		0.521	0.552	0.575	0.597	0.611	0.523	0.553	0.575	0.598	0.611
GradientBoostingRegressor		0.686	0.691	0.695	0.694	0.691	0.689	0.689	0.698	0.691	0.694
ExtraTreesRegressor		0.465	0.501	0.531	0.559	0.588	0.467	0.499	0.532	0.56	0.587
XGBRegressor		0.664	0.641	0.66	0.642	0.653	0.664	0.641	0.66	0.642	0.653
CatBoostRegressor		0.694	0.701	0.709	0.706	0.702	0.694	0.704	**0.716**	0.714	0.708

m^3, respectively. The wind direction (WD) in summer, with the highest T (28.83 °C), RH (74.21%), and Ox (112.29 μg/m^3) was mainly south and southeast. In spring, with the strongest SSR (629.68 kW/m^2) and slightly lower Ox than that in summer (106.31 μg/m^3), southerly winds prevailed. The dominant WD in winter, with the lowest T (2.62 °C), was northeast and northwest.

The annual average concentrations of the 30 VOC components in the PMF source apportionment for Huaibei in 2022 are listed in Table S4. The three components with the highest average concentrations were ethane, propane, and acetone, with concentrations of 3.336, 2.348, and 2.406 ppb, respectively. The 30 VOC components were divided into six categories: alkanes, alkenes, alkynes, OVOCs, aromatics, and halohydrocarbons (Fig. S3). Among the six categories, alkanes (50.2%) constituted the highest proportion of TVOCs, followed by OVOCs (16.2%). Fig. S3 showed that the concentrations of TVOCs exhibit an increase and decrease in winter and summer, respectively. The decrease in summer has been attributed to a greater boundary layer height and rapid photochemical removal (Goldstein et al., 1995; Kramer et al., 2015).

3.2. Source apportionment of VOCs

In this study, the PMF model was used to quantify the contributions of sources. The number of sources was set from four to nine, and the results of seven sources were selected while ensuring adequate fit to the measurement data and the best physical meaning (i.e., details are provided in Table S5). The factor rotating results were not significantly different from those of the rotation results. Thus, the results used in this study were from the runs with Fpeak = 0.0, which provided the most physically reasonable source profiles (shown in Fig. 3). Seven sources, namely, industrial sources (IS), oil and gas volatilization sources (OGVS), vehicle exhaust (VE), combustion sources (COS), biological sources (BS), catering sources (CS) and solvent sources (SS), were identified using the PMF mode. The source profiles are shown in Fig. 3.

The major species in Factor 1 were low-carbon alkanes and halohydrocarbons (dichloromethane; methyl chloride; and 1,2-dichloroethane), with higher proportions of benzene, toluene, and ethylbenzene. Of these, 1,2-dichloroethane is an intermediate product of vinyl chloride production in the petrochemical industry (Liu et al., 2023a; Wang et al., 2022b). Benzene may also originate from industrial processes (Cai et al., 2010). Therefore, Factor 1 was determined to be IS.

The source profile of Factor 2 was characterized by high proportions of iso/n-butane and iso/n-pentane, which are derived from OGVS (Liu et al., 2023a). Thus, Factor 2 was determined to be the OGVS.

Factor 3 was characterized by relatively high proportions of C4–C6 alkanes and acetylene (39.1%), which are important indicators of VE (Song et al., 2021).

Factor 4 had a high percentage of ethane (41.2%), propane (37.3%), and ethylene (33.9%), which conformed to the emission characteristics of COS (Ling et al., 2011; Song et al., 2021). Factor 5, identified as BS, was characterized by a high percentage of isoprene. Isoprene is a tracer of biogenic emissions (Wang et al., 2022b), accounting for 71.4% of the total emission factors.

The proportion of ethyl acetate (71.8%) in Factor 6 was high, indicating that Factor 6 was CS (Chen et al., 2022; Klein et al., 2016).

Factor 7 was rich in aromatic compounds such as m,p-xylene (74.0%), o-xylene (63.5%), toluene (52.3%), and ethylbenzene (45.4%). A high proportion of these species mainly originate from the SS (An et al., 2017; Li et al., 2018).

Fig. 4 shows the source appointment results of PMF, wherein the contribution of each source to the TVOCs during the observation period was calculated. VE and COS were the dominant sources of VOCs with contributions of 25.0% and 22.1%, respectively. IS, OGVS, SS, CS, and BS also contributed to the VOCs, accounting for 16.6%, 15.9%, 9.5%, 8.5%, and 2.5%, respectively. Regarding seasonal factors, VE (26.0%) had the highest contribution to VOCs in spring, followed by industrial sources (22.3%), oil and gas volatilization sources (16.6%), and combustion sources (15.3%), with other emission sources contributing

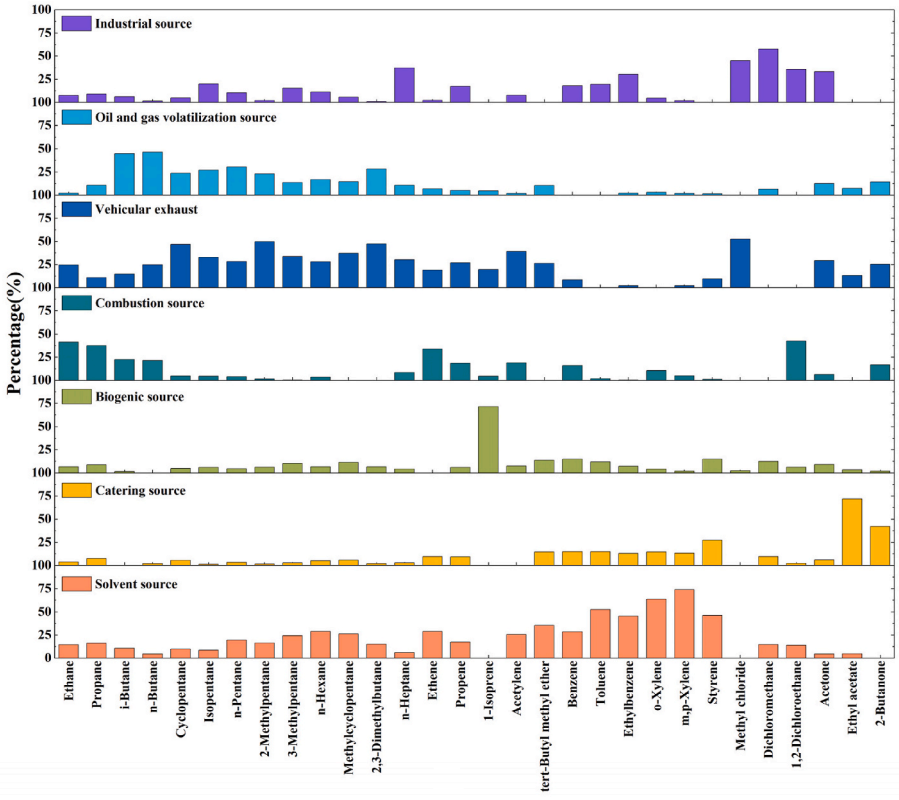


Fig. 3. Factor spectrum results of PMF.

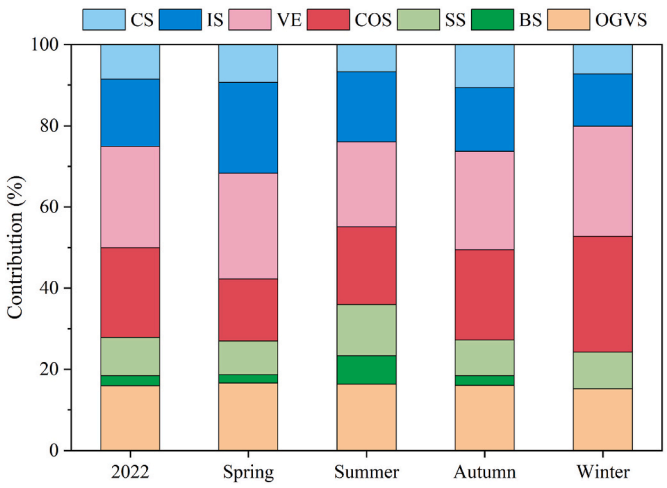


Fig. 4. Analysis of VOCs sources of Huaibei in 2022.

between 2.0% and 9.4%. The main sources of VOCs in the summer were VE (21.0%), COS (19.1%), IS (17.2%), OGVS (16.3%), and SS (12.6%). VE (24.2%) and COS (22.2%) were the main sources of VOCs in autumn, while COS (28.5%), VE (27.1%), OGVS (15.1%), and IS (12.9%) were the main sources of VOCs in winter. The contribution of COS to VOCs in autumn and winter was higher than that in spring and summer, with the contribution of COS in winter approximately twice that in summer, possibly due to heating in the north during winter (Guan et al., 2023), which leads to an increase in the contribution of COS from pollution transport (Song et al., 2021).

The proportion of VE was the highest in all seasons (>20%), whereas the proportion of natural sources was the lowest in all seasons, ranging from 0 to 7.0%. Overall, VE, COS, and IS emissions contributed significantly to the concentration of VOCs in Huaibei. Our findings are corroborated by several studies that have shown that VE is the main source of VOCs in large cities, such as the Pearl River Delta (He et al., 2019), Xi'an (Song et al., 2021) and on summer days in Beijing (Li et al., 2015). Moreover, VE has been shown to account for a large proportion of VOCs in Shanghai's winter (Liu et al., 2021; Wang et al., 2022b). However, there are some differences in specific seasons or regions, such as the higher contribution of COS to the winter heating period in Beijing (Niu et al., 2022), and the most significant impact of IS and VE on VOCs in Shijiazhuang (Guan et al., 2023).

3.3. Sources and meteorological influences on VOCs pollution identified through ML

3.3.1. Feature importance analysis

The CatBoost regression model was trained and evaluated using datasets for the entire year of 2022, including spring (March to May), summer (June to August), autumn (September to November), and winter (January, February, and December), with R^2 as the evaluation metric. As shown in Table S6, the R^2 of *CatBoostRegression* models trained on datasets for all periods were higher than 0.77, indicating a good fitting performance.

The proportion of feature importance of each driver factor was obtained using the CatBoost regression model (Fig. 5), which showed that T and RH had a higher influence on TVOC concentrations throughout 2022 compared with other factors. The dominant impact of T is mainly due to it being fully representative of the seasonal variation of pollutant concentrations (Zhang et al., 2022). Additionally, the secondary transformation of VOCs to O_3 is more conducive at high T and low RH (Liu et al., 2022b, 2023a, 2023b; Song et al., 2021). Across the four seasons, T had a lower importance, whereas RH had a higher importance, on TVOC concentrations in spring, autumn, and winter. RH had the greatest effect on TVOC concentrations in spring, VE had the most influence in summer, CS had the greatest influence in autumn, and SS had the highest influence on the concentration of TVOCs in winter.

3.3.2. Driving factor analysis

The proportion of feature importance based on ML models can describe the overall impact of various driving factors on TVOCs; however, obtaining the quantitative impact of driving factors on model predictions is impossible. Therefore, the SHAP model was used to analyze the CatBoost regression model, and the impact of all driving factors was obtained using the constructed CatBoost-SHAP method. The SHAP values displayed fluctuations in each predicted value under various driving factors. Moreover, the SHAP model could promote or reduce the concentration of TVOCs through various driving factors at different time periods by explaining the ML training results.

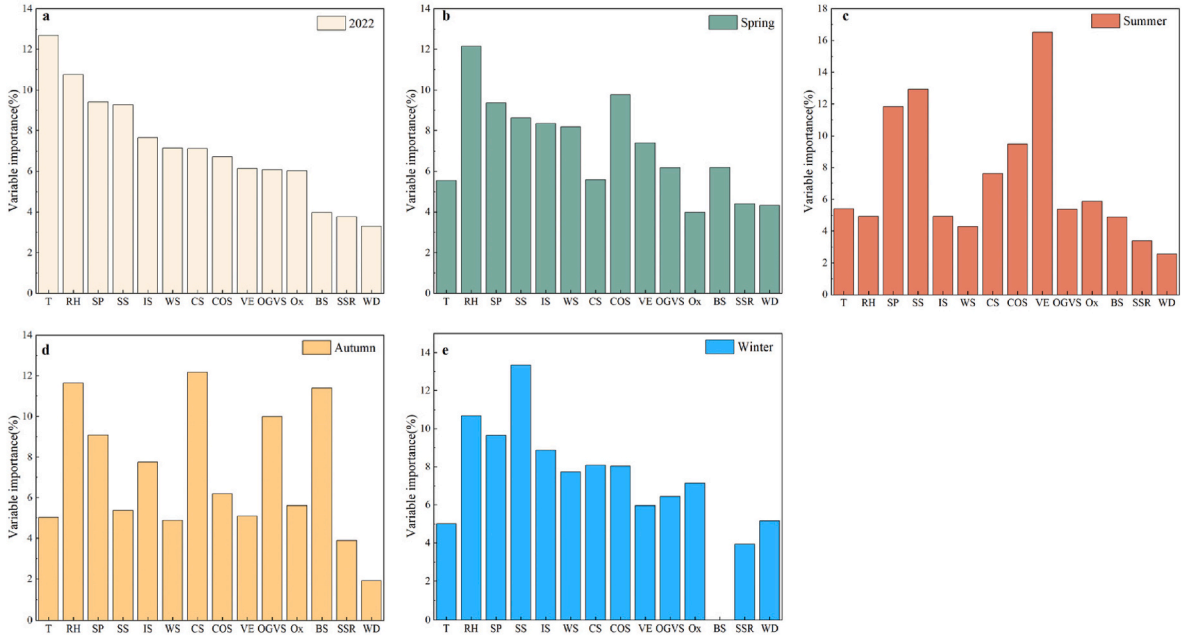


Fig. 5. Feature importance of TVOCs driving factors in Huaibei for 2022 and four seasons obtained through CatBoostRegression model.

The benchmark values of the SHAP model for the entire year of 2022, as well as for spring, summer, autumn, and winter, were 18.7, 17.2, 14.5, 20.4, and 22.5 ppb, respectively. Fig. 6 shows the proportions of driving factors for TVOCs in Huaibei in 2022 under CatBoost-SHAP. Meteorological and emission factors had relatively large impacts on TVOCs, accounting for 50.9% and 42.5%, respectively, whereas atmospheric oxidation accounted for only 6.6%. T and RH had relatively large meteorological impacts, accounting for 30.8% and 28.8% of the variance, respectively. This conclusion is consistent with that of the characteristically important proportion of VOC-driving factors in 2022. Regarding the influence of emissions, SS and BS had greater impacts, accounting for 24.7% and 21.1%, respectively. Regarding the impact of atmospheric oxidation, Ox accounted for 63.3% and SSR accounted for 36.7% of the variation.

The SHAP values and proportion of SHAP values of the driving factors are shown in Figs. 7 and **8**, respectively. From an annual perspective, T was the largest driving factor affecting the TVOC concentrations, followed by RH. Seasonally, the impact of T decreased from 1.6 ppb to 0.27, 0.23, 0.54 and 0.47 ppb in spring, summer, autumn and winter, with the decline ratio of 83.6%, 85.7%, 66.4% and 70.7%, respectively. RH had a significant impact throughout the year, spring, autumn, and winter, whereas its impact on summer was small. Meanwhile, SS had a low influence on the concentration of VOCs throughout the year (spring, summer, and winter), whereas it had the highest impact during winter (>2 ppb). The impact of VE on TVOCs throughout the year (spring, autumn, and winter) was less than 1 ppb; however, it had the highest impact in summer.

Based on Figs. 7 and 8, the influences of meteorology on spring, summer, autumn, and winter were 37.8%, 24.6%, 38.3%, and 40.9%, respectively, and the impacts of the emissions during spring, summer, autumn, and winter were 56.0%, 64.5%, 51.8%, and 50.4%, respectively. Of the meteorological elements, RH had the greatest impact on TVOCs in spring and autumn 2022 accounting for 41.7% and 44.3%, respectively; VE had the greatest impact on summer TVOCs (32.4%), and SS had the greatest impact on winter TVOCs (33.7%).

3.4. In-depth analysis of the summer period

To determine the impact of the driving factors on TVOCs concentrations during the O_3 pollution and non-pollution periods in Huaibei, two typical time intervals, from June 15th to June 19th and June 25th to June 28th, were selected. Based on the CatBoost-SHAP model and SHAP values, the quantitative impacts of all driving factors on TVOCs during these typical periods were analyzed. Table S7 shows the daily average of O_3 during various time periods; Fig. 9 shows that O_3 in Huaibei exhibited obvious diurnal variation characteristics, reaching its highest value from noon to the afternoon.

Fig. 9 shows the SHAP hourly values of the driving factors of TVOCs during typical periods in Huaibei in 2022, and the distribution of SHAP values for all driving factors. Overall, during the period of ozone pollution, the SHAP value of Ox increased owing to the increase in radiation intensity and T, peaking at noon (Fig. 9 and Fig. S4) (Fu et al., 2020a; Hou et al., 2022). The SHAP value of the daytime SSR increased earlier than that of Ox, possibly due to photochemical reactions occurring during the day (Hou et al., 2022). The SHAP values of RH and T during the pollution and non-pollution periods showed a similar pattern, with a negative impact on the TVOCs concentrations during the day and a positive impact at night. This result illustrates that a high RH and low T are beneficial for the accumulation of TVOCs. Specifically, when the pollution level was severe (O_3-8h: 220 μg/m^3), the SHAP values of all driving factors were positive, indicating that all driving factors promoted an increase in the concentration of TVOCs. Moreover, as the SHAP values of the driving factors gradually decreased, O_3 concentrations gradually peaked. Based on Fig. 9 (b) and (d), from June 15th to June 19th and from June 25th to June 28th, the contribution of SS to TVOCs was positive, indicating the need for stronger control over solvent sources.

During the two typical periods, emissions were the main cause of changes in TVOCs concentrations (Fig. 10). Under polluted conditions, emissions accounted for 64.3% of the total variance, with SS accounting for 22.5%. The impact of meteorology accounted for 23.3%, with SP having the greatest impact (6.8%), followed by T (5.2%). Atmospheric oxidation accounted for 12.4%, with Ox accounting for 8.0%, and SSR accounting for 4.4%. During the non-polluting periods, emissions

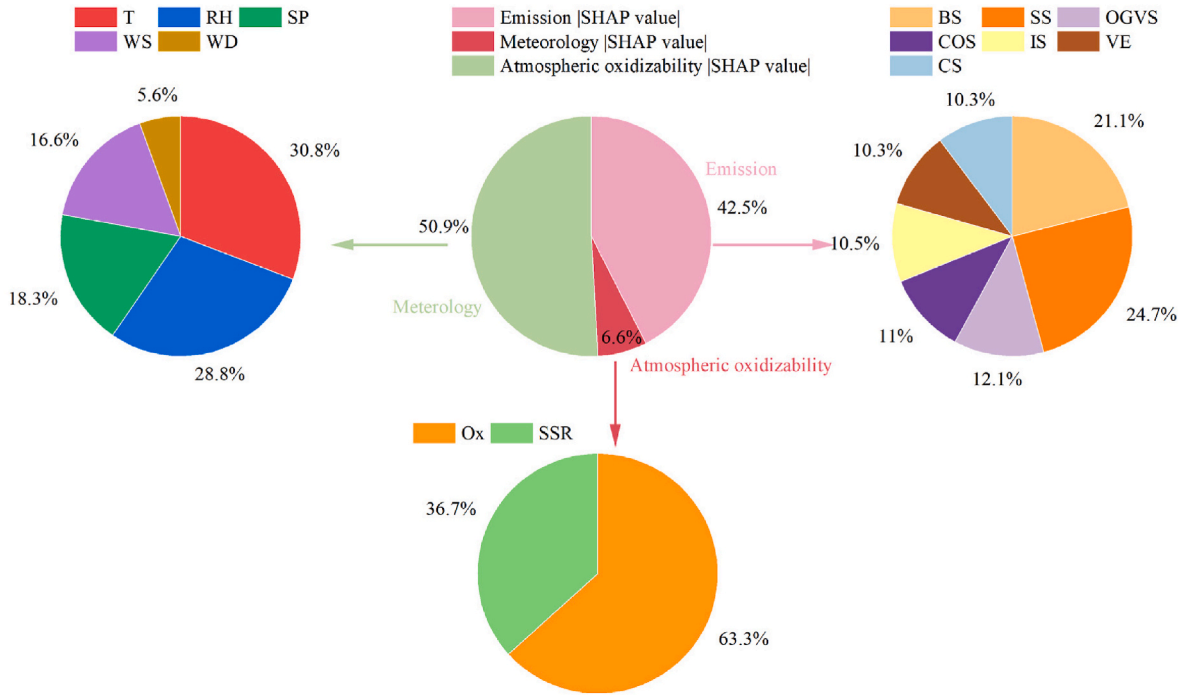


Fig. 6. The proportion of factors that shaped TVOCs in Huaibei in 2022, as evaluated through the CatBoost-SHAP model.

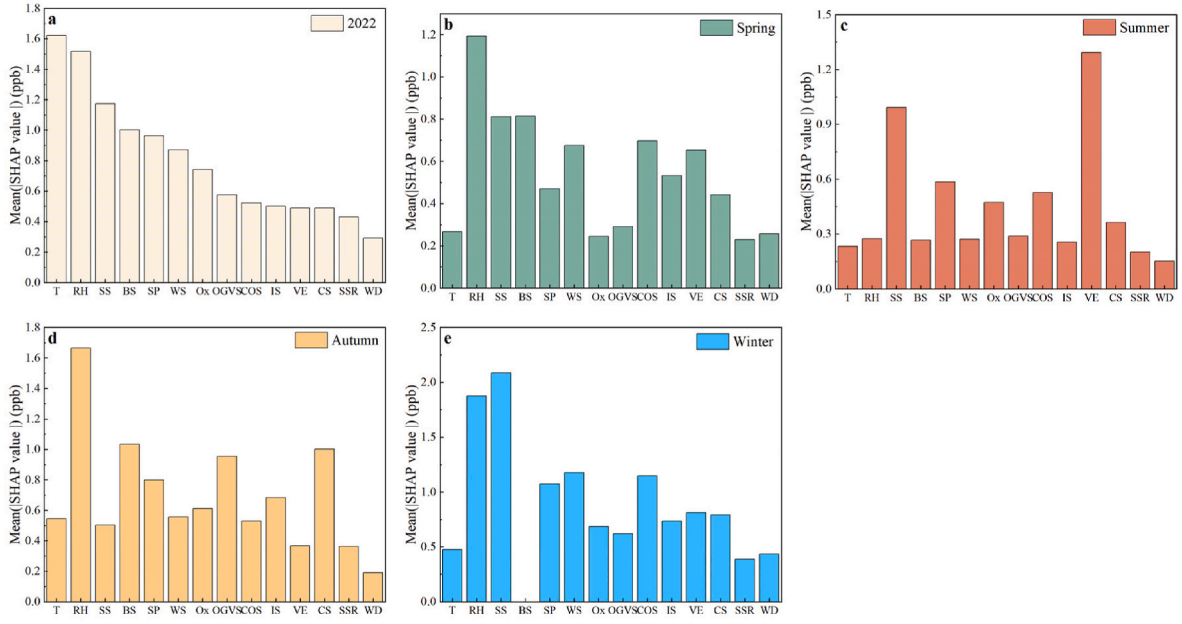


Fig. 7. The SHAP values of factors that driving TVOCs in Huaibei.

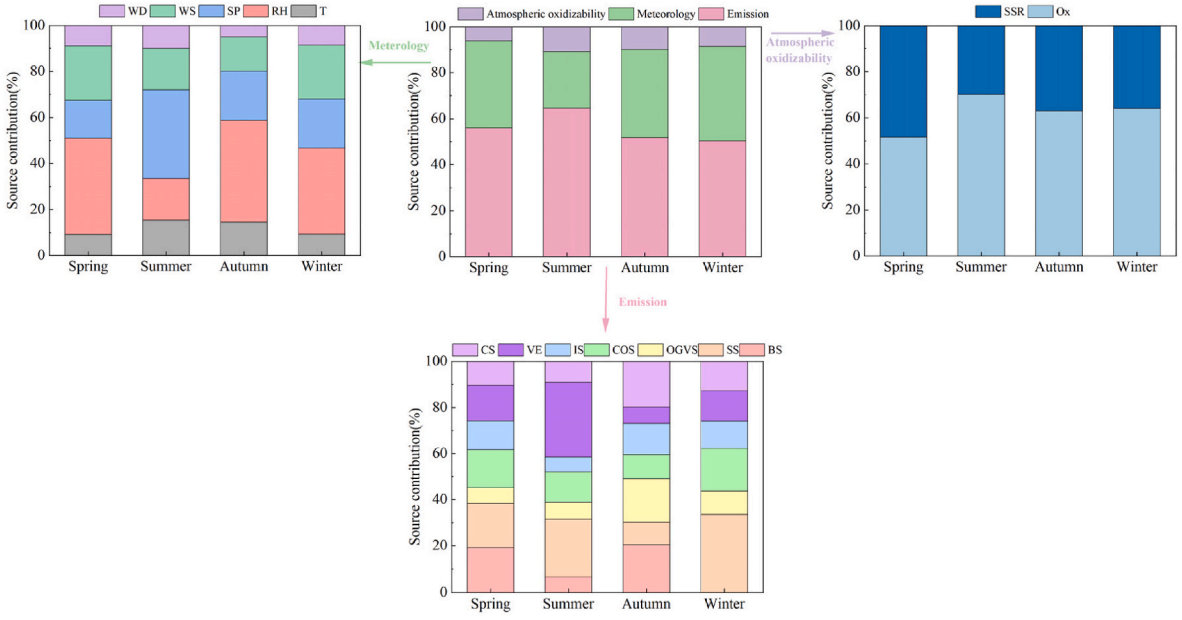


Fig. 8. The contribution of SHAP values of factors driving TVOCs in Huaibei.

accounted for 63.7% of the total, with VE having the greatest impact at 20.7%. The impact of meteorology accounted for 26.0%, with SP having the greatest impact (8.7%), followed by WS (6.9%). Finally, atmospheric oxidation accounted for 10.2%, with Ox accounting for 7.6%, and SSR accounting for 2.6%.

4. Conclusions

VOCs are important precursors of O_3 and $PM_{2.5}$. Therefore, quantifying the impact of these factors has become a focal point in the field of pollution control. The complex nonlinear relationship between TVOCs and their driving factors (i.e., emissions, meteorological factors, and atmospheric oxidation) makes it difficult to evaluate the impact of each driving factor on VOCs concentrations using conventional methods. In this study, a machine learning method (CatBoost-SHAP) combined with the PMF model was applied to explore the impacts of driving factors on VOC concentrations based on a one-year dataset (2022) representative of Huaibei, China. The insights brought forth characterize the impact of various factors on VOC concentrations across time. During the sampling period, the concentration of VOCs in the ambient air showed high seasonal variations in winter and low seasonal variations in summer. Seven VOC sources were extracted using the PMF model: IS, OGVS, VE, COS, BS, CS, and SS. VE, COS, and IS contributed significantly to the concentration of VOCs. The contributions of emission sources varied in different seasons.

The results of the CatBoost-SHAP model indicated that approximately 42.5% of the variation in VOC concentration was attributed to the impact of emissions, 50.9% was meteorological, and only 6.6% was

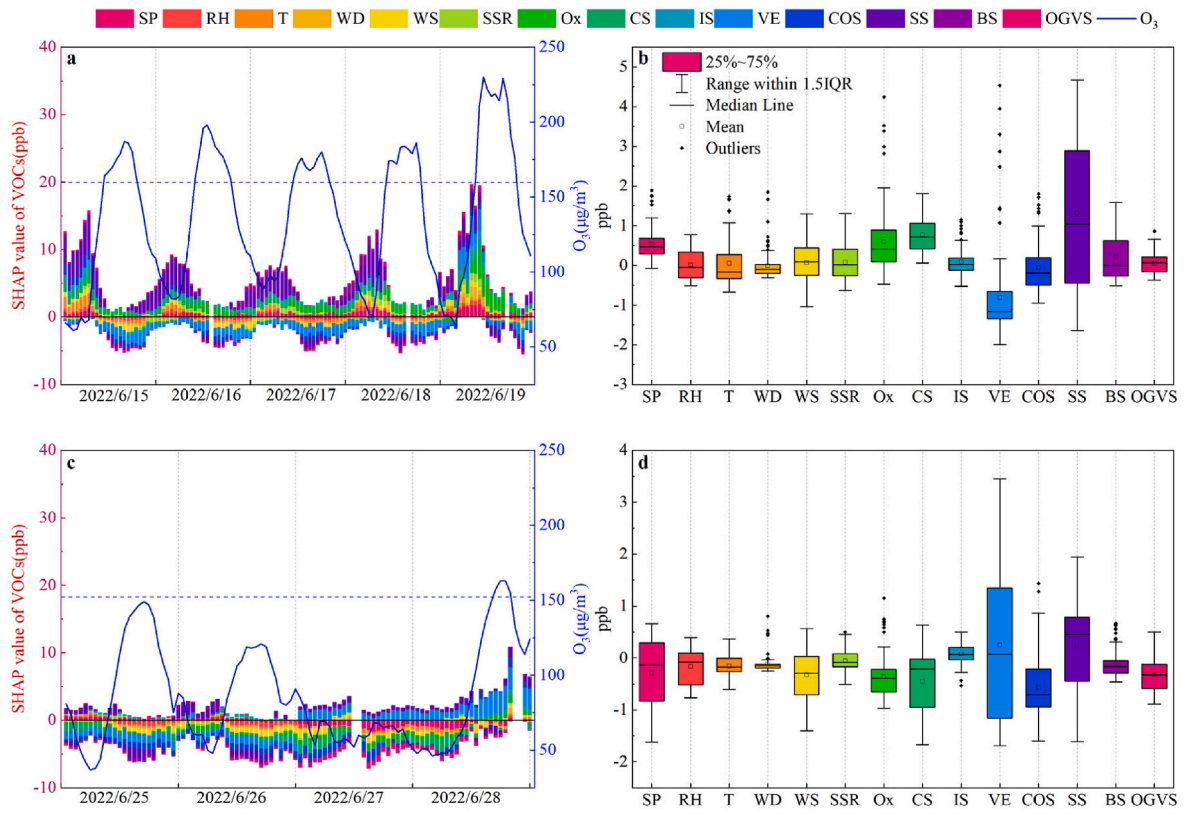


Fig. 9. SHAP values of driving factors for TVOCs during specific periods in 2022.

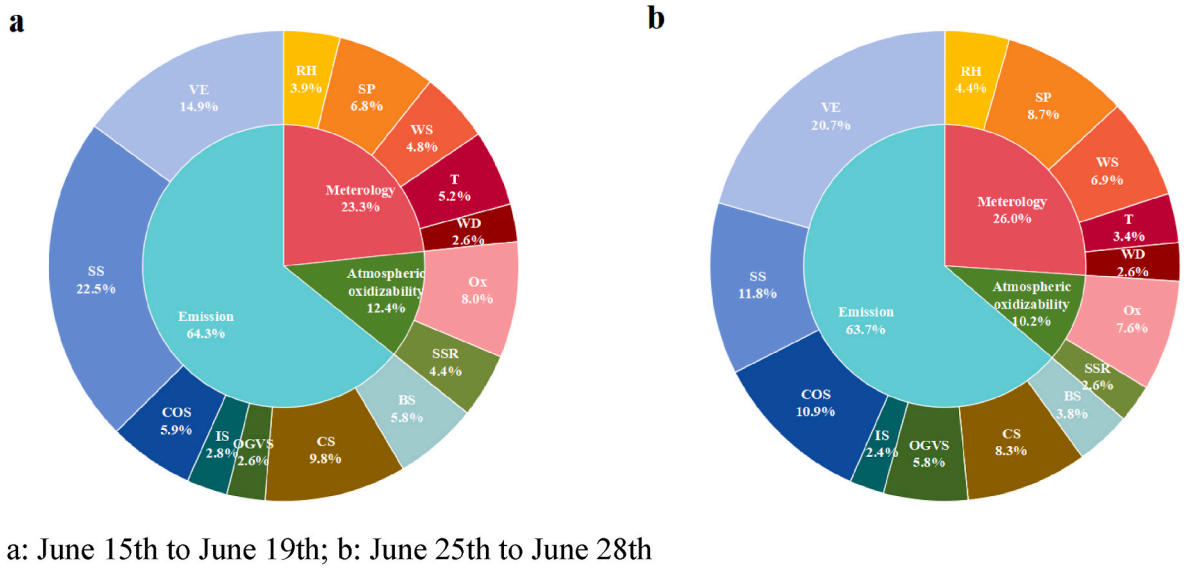


Fig. 10. SHAP value proportion of driving factors for TVOCs.

atmospheric oxidative. Our results suggest that VE and SS are the dominant sources of VOCs. Thus, a reduction in VE and SS can effectively decrease the levels of VOCs and alleviate O_3 pollution.

In this study, a machine learning model (CatBoost-SHAP) was inputted with hourly resolution datasets and VOC source apportionment results, revealing the quantitative impact of various driving factors on VOCs in different scenarios. Our findings bring forth insights that can be channeled toward the development of effective air quality management strategies, since they identify major pollution sources, thus supporting targeted governance. Moreover, this study will help formulate effective air quality management strategies by identifying ways to control primary emissions that can effectively reduce VOC emissions, which could provide references for local governments regarding collaborative strategies for controlling $PM_{2.5}$ and O_3.

CRediT authorship contribution statement

Wei Chen: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Xuezhe Xu:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources,

Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Wenqing Liu:** Writing – review & editing, Visualization, Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Research Program for Key Issues in Air Pollution Control (No. DQGG202117), and the Key Research and Development Plan of Anhui Province (No. 2023t07020009).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2024.120714.

References

An, J.L., Wang, J.X., Zhang, Y.X., Zhu, B., 2017. Source apportionment of volatile organic compounds in an urban environment at the yangtze River Delta, China. Arch. Environ. Con. Tox. 72, 335–348. https://doi.org/10.1007/s00244-017-0371-3.

Bo, Y., Liu, Q.S., Huang, S., Pan, Y.C., 2022. Real-time hard-rock tunnel prediction model for rock mass classification using CatBoost integrated with Sequential Model-Based Optimization. Tunn. Uudergr. Sp. Tech. 124, 104448 https://doi.org/10.1016/j.tust.2022.104448.

Cai, C.J., Geng, F.H., Tie, X.X., Yu, Q., An, J.L., 2010. Characteristics and source apportionment of VOCs measured in Shanghai, China. Atmos. Environ. 44 (38), 5005–5014. https://doi.org/10.1016/j.atmosenv.2010.07.059.

Chen, P., Zhang, Y., Xing, M., Li, S.S., 2022. Classification control of volatile organic compounds (VOCs) emission pollution sources based on emission amounts and atmospheric reactivity. Environ. Sci. 43 (5), 2383–2394. https://doi.org/10.13227/j.hjkx.202108204.

Cheng, J., Zhang, Y.S., Wang, T., Xu, H., Norris, P., Pan, W.P., 2018. Emission of volatileorganic compounds (VOCs) during coal combustion at different heating rates. Fuel 225, 554–562. https://doi.org/10.1016/j.fuel.2018.03.185.

Dai, Q.L., Ding, J., Song, C.B., Liu, B.S., Bi, X.H., Wu, J.H., et al., 2021. Changes in source contributions to particle number concentrations after the COVID-19 outbreak: insights from a dispersion normalized PMF. Sci. Total Environ. 759, 143548 https://doi.org/10.1016/j.scitotenv.2020.143548.

Dai, H.B., Huang, G.Q., Wang, J.J., Zeng, H.B., 2023. VAR-tree model based spatio-temporal characterization and prediction of O_3 concentration in China. Ecotox. Environ. Safe. 257, 114960 https://doi.org/10.1016/j.ecoenv.2023.114960.

Drozd, G.T., Zhao, Y.L., Saliba, G., Frodin, B., Maddox, C., Weber, R.J., et al., 2016. Time resolved measurements of speciated tailpipe emissions from motor vehicles: trends with emission control technology, cold start effects, and speciation. Environ. Sci. Technol. 50, 13592–13599. https://doi.org/10.1021/acs.est.6b04513.

Fu, X., Wang, T., Gao, J., Wang, P., Liu, Y.M., Wang, S.X., et al., 2020a. Persistent heavy winter nitrate pollution driven by increased photochemical oxidants in northern China. Environ. Sci. Technol. 54 (7), 3881–3889. https://doi.org/10.1021/acs.est.9b07248.

Fu, S., Guo, M.X., Luo, J.M., Han, D.M., Chen, X.J., Jia, H.H., et al., 2020b. Improving VOCs controlstrategies based on source characteristics and chemical reactivity in a typical coastal city of South China through measurement and emission inventory. Sci. Total Environ. 744, 140825 https://doi.org/10.1016/j.scitotenv.2020.140825.

Gao, J., Dong, S.H., Yu, H.F., Peng, X., Wang, W., Shi, G.L., et al., 2020. Source apportionment for online dataset at a megacity in China using a new PTT-PMF model. Atmos. Environ. 229, 117457 https://doi.org/10.1016/j.atmosenv.2020.117457.

Goldstein, A.H., Wofsy, S.C., Spivakovsky, C.M., 1995. Seasonal variations of nonmethane hydrocarbons in rural New England: constraints on OH concentrations in northern midlatitudes. J. Geophys. Res. 100 (D10) https://doi.org/10.1029/95jd03428, 21023-21033.

Guan, Y.N., Liu, X.J., Zheng, Z.Y., Dai, Y.W., Du, G.M., Han, J., et al., 2023. Summer O_3 pollution cycle characteristics and VOCs sources in a central city of Beijing-Tianjin-Hebei area, China. Environ. Pollut. 323 https://doi.org/10.1016/j.envpol.2023.121293, 121293-121293.

Guo, R.H., Zhu, X.F., Zhu, Z.G., Sun, J.H., Li, Y.Z., Hu, W.C., et al., 2022. Evaluation of typical volatile organic compounds levels in new vehicles under static and driving conditions. Int. J. Environ. Res. Publ. Health 19 (12), 7048. https://doi.org/10.3390/ijerph19127048.

Hancock, J., Khoshgoftaar, T.M., 2020. CatBoost for big data: an interdisciplinary review. Journal of Big Data 7 (1), 1–45. https://doi.org/10.1186/s40537-020-00369-8.

Hang, J., Wang, X.M., Liang, J., Zhang, X.L., Wu, L.L., Du, Y.X., et al., 2023. Numerical investigation of the impact of urban trees on O_3-NOx-VOCs chemistry and pollutant dispersion in a typical street canyon. Atmos. Environ. 311, 119998 https://doi.org/10.1016/j.atmosenv.2023.119998.

He, Z.R., Wang, X.M., Ling, Z.H., Zhao, J., Guo, H., Shao, M., et al., 2019. Contributions of different anthropogenic volatile organic compound sources to ozone formation at a receptor site in the Pearl River Delta region and its policy implications. Atmos. Chem. Phys. 19 (13), 8801–8816. https://doi.org/10.5194/acp-19-8801-2019.

Hou, L.L., Dai, Q.L., Song, C.B., Liu, B.W., Guo, F.Z., Dai, T.J., et al., 2022. Revealing drivers of haze pollution by explainable machine learning. Environ. Sci. Technol. Lett. 9 (2), 112–119. https://doi.org/10.1021/acs.estlett.1c00865.

Hu, C.Y., Kang, P., Jaffe, D.A., Li, C.K., Zhang, X.L., Wu, K., et al., 2021. Understanding the impact of meteorology on ozone in 334 cities of China. Atmos. Environ. 248, 118221 https://doi.org/10.1016/j.atmosenv.2021.118221.

Huang, G.M., Wu, L.F., Ma, X., Zhang, W.Q., Fan, J.L., Yu, X., et al., 2019. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J. Hydrol. 574, 1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085.

Hunter-Sellars, E., Tee, J.J., Parkin, I.P., Williams, D.R., 2020. Adsorption of volatile organic compounds by industrial porous materials: impact of relative humidity. Micropor. Mesopor. Mat. 298, 110090 https://doi.org/10.1016/j.micromeso.2020.110090.

Jabeur, S.B., Gharib, C., Mefteh-Wali, S., Arfi, W.B., 2021. CatBoost model and artificial intelligence techniques for corporate failure prediction. Technol. Forecast. Soc. Change 166, 120658. https://doi.org/10.1016/j.techfore.2021.120658.

Jia, Y.C., Andersen, H., Cermak, 2023. Analysis of Cloud Fraction Adjustment to Aerosols and its Dependence on Meteorological Controls Using Explainable Machine Learning. EGUsphere. https://doi.org/10.5194/egusphere-2023-1667.

Klein, F., Platt, S.M., Farren, N.J., Detournay, A., Bruns, E.A., Bozzetti, C., et al., 2016. Characterization of gas-phase organics using proton transfer reaction yime-of-flight mass spectrometry: cooking emissions. Environ. Sci. Technol. 50 (3), 1243–1250. https://doi.org/10.1021/acs.est.7b03960.

Kramer, L.J., Helmig, D., Burkhart, F., Stohl, A., Oltmans, S., Honrath, R.E., 2015. Seasonal variability of atmospheric nitrogen oxides and non-methane hydrocarbonsat the GEOSummit station, Greenland. Atmos. Chem. Phys. 15, 6827–6849. https://doi.org/10.5194/acp-15-6827-2015.

Li, L.Y., Xie, S.D., Zeng, L.M., Wu, R.R., Li, J., 2015. Characteristics of volatile organic compounds and their role in ground-level ozone formation in the Beijing-Tianjin-Hebei region, China. Atmos. Environ. 113, 247–254. https://doi.org/10.1016/j.atmosenv.2015.05.021.

Li, J., Zhai, C.Z., Yu, J.Y., Liu, R.L., Li, Y.Q., Zeng, L.M., et al., 2018. Spatiotemporal variations of ambient volatile organic compounds and their sources in Chongqing, a mountainous megacity in China. Sci. Total Environ. 627, 1442–1452. https://doi.org/10.1016/j.scitotenv.2018.02.010.

Li, K., Jacob, D.J., Liao, H., Shen, L., Zhang, Q., Bates, K.H., 2019. Anthropogenic drivers of 2013-2017 trends in summer surface ozone in China. Proc. Natl. Acad. Sci. USA 116 (2), 422–427. https://doi.org/10.1073/pnas.1812168116.

Li, K., Jacob, D.J., Shen, L., Lu, X., De Smedt, I., Liao, H., 2020. Increases in surface ozone pollution in China from 2013 to 2019: anthropogenic and meteorological influences. Atmos. Chem. Phys. 20 (19), 11423–11433. https://doi.org/10.5194/acp-20-11423-2020.

Liang, Q., Bao, X., Sun, Q., Zhang, Q.L., Zou, X., Huang, C.Q., et al., 2020. Imaging VOC distribution in cities and tracing VOC emission sources with a novel mobile proton transfer reaction mass spectrometer. Environ. Pollut. 265, 114628 https://doi.org/10.1016/j.envpol.2020.114628.

Ling, Z.H., Guo, H., Cheng, H.R., Yu, Y.F., 2011. Sources of ambient volatile organic compounds and their contributions to photochemical ozone formation at a site in the Pearl River Delta, southern China. Environ. Pollut. 159, 2310–2319. https://doi.org/10.1016/j.envpol.2011.05.001.

Ling, Z.H., Guo, H., 2014. Contribution of VOC sources to photochemical ozone formation and its control policy implication in Hong Kong. Environ. Sci. Pol. 38, 180–191. https://doi.org/10.1016/j.envsci.2013.12.004.

Liu, H., Man, H.Y., Tschantz, M., Wu, Y., He, K.B., Hao, J.M., 2015. VOC emissions from the vehicle evaporation process: status and control strategy. Environ. Sci. Technol. 49, 14424–14431. https://doi.org/10.1021/acs.est.5b04064.

Liu, C.Q., Shi, K., 2021. A review on methodology in O_3-NOx-VOC sensitivity study. Environ. Pollut. 291, 118249 https://doi.org/10.1016/j.envpol.2021.118249.

Liu, Y.H., Wang, H.L., Jing, S.G., Peng, Y.R., Gao, Y.Q., Yan, R.S., et al., 2021. Strong regional transport of volatile organic compounds (VOCs) during wintertime in Shanghai megacity of China. Atmos. Environ. 244, 117940 https://doi.org/10.1016/j.atmosenv.2020.117940.

Liu, X., Lu, D., Zhang, A., Liu, Q., Jiang, G., 2022a. Data-Driven machine learning in environmental pollution: gains and problems. Environ. Sci. Technol. 56, 2124–2133. https://doi.org/10.1021/acs.est.1c06157.

Liu, Y.F., Qiu, P.P., Li, C.L., Li, X.K., Ma, W., Yin, S.J., et al., 2022b. Evolution and variations of atmospheric VOCs and O_3 photochemistry during a summer O_3 event in a county-level city, Southern China. Atmos. Environ. 272, 118942 https://doi.org/10.1016/j.atmosenv.2022.118942.

Liu, T.H., Cheng, L., Wang, C.D., Wang, R.P., Yao, B.B., Wei, W., 2023a. Research on VOCs sources and local O_3 generation in typical industrial cities. Acta Sci. Circumstantiae. https://doi.org/10.13671/j.hjkxxb.2023.0246, 2023.

Liu, Z.G., Wang, B.L., Wang, C., Sun, Y.C., Zhu, C.Y., Sun, L., et al., 2023b. Characterization of photochemical losses of volatile organic compounds and their implications for ozone formation potential and source apportionment during summer in suburban Jinan, China. Environ. Res. 238 (1), 117158 https://doi.org/10.1016/j.envres.2023.117158.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long, Beach, CA, USA. https://doi.org/10.48550/arXiv.1705.07874. Dec 4.

Markowicz, P., Larsson, L., 2015. Influence of relative humidity on VOC concentrations in indoor air. Environ. Sci. Pollut. R. 22, 5772–5779. https://doi.org/10.1007/s11356-014-3678-x.

Mo, Z.W., Shao, M., Lu, S.H., Niu, H., Zhou, M.Y., Sun, J., 2017. Characterization of non-methane hydrocarbons and their sources in an industrialized coastal city, Yangtze River Delta. China. Sci. Total Environ. 593–594, 641–653. https://doi.org/10.1016/j.scitotenv.2017.03.123.

Mozaffar, A., Zhang, Y.L., 2020. Atmospheric volatile organic compounds (VOCs) in China: a review. Curr. Pollut. Rep. 6 (3), 250–263. https://doi.org/10.1007/s40726-020-00149-1.

Nelson, D., Choi, Y., Sadeghi, B., Yeganeh, A.K., Ghahremanloo, M., Park, J., 2023. A comprehensive approach combining positive matrix factorization modeling, meteorology, and machine learning for source apportionment of surface ozone precursors: underlying factors contributing to ozone formation in Houston, Texas. Environ. Pollut. 334, 122223 https://doi.org/10.1016/j.envpol.2023.122223.

Niu, Y.Y., Yan, Y.L., Chai, J.W., Zhang, X.Y., Xu, Y., Duan, X.L., et al., 2022. Effects of regional transport from different potential pollutionareas on volatile organic compounds (VOCs) in Northern Beijing during non-heatingand heating periods. Sci. Total Environ. 836, 155465 https://doi.org/10.1016/j.scitotenv.2022.155465.

Nussbaumer, C.M., Cohen, R.C., 2020. The role of temperature and NOx in ozone trends in the Los Angeles Basin. Environ. Sci. Technol. 54 (24), 15652–15659. https://doi.org/10.1021/acs.est.0c04910.

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemometr. Intell. Lab. Syst. 37 (1), 23–35. https://doi.org/10.1016/S0169-7439(96)00044-5.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126. https://doi.org/10.1002/env.3170050203.

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 (4), 2189–2207. https://doi.org/10.1021/acs.est.0c07204.

Peng, ZeZ., Zhang, B., Wang, D.W., Niu, X.Y., Sun, J., Xu, H.G., et al., 2023. Application of machine learning in atmospheric pollution research: a state-of-art review. Sci. Total Environ. 910, 168588 https://doi.org/10.1016/j.scitotenv.2023.168588.

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. https://doi.org/10.1016/j.jes.2020.10.005.

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., Sonne, C., et al., 2023. Air quality prediction by machine learning models: a predictive study on the indian coastal city of Visakhapatnam. Chemosphere 338, 139518. https://doi.org/10.1016/j.chemosphere.2023.139518.

Reid, C., Jerrett, M., Petersen, M., Pfister, G., Morefield, P., Tagar, I.B., et al., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ. Sci. Technol. 49 (6), 3887–3896. https://doi.org/10.1021/es505846r.

Sadeghi, B., Ghahremanloo, M., Mousavinezhad, S., Lops, Y., Pouyaei, A., Choi, Y., 2022. Contributions of meteorology to ozone variations: application of deep learning and the Kolmogorov-Zurbenko filter. Environ. Pollut. 310, 119863 https://doi.org/10.1016/j.envpol.2022.119863.

Sayeed, A., Choi, Y., Pouyaei, A., Lops, Y., Jung, J., Salman, A.K., 2022. CNN-based model for the spatial imputation (CMSI version 1.0) of in-situ ozone and $PM_{2.5}$ measurements. Atmos. Environ. 289, 119348 https://doi.org/10.1016/j.atmosenv.2022.119348.

Song, M.D., Li, X., Yang, S.D., Yu, X.N., Zhou, S.X., Yang, Y.M., et al., 2021. Spatio-temporal variation, sources, and secondary transformation potential of volatile organic compounds in Xi'an, China. Atmos. Chem. Phys. 21, 4939–4958. https://doi.org/10.5194/acp-21-4939-2021.

Tan, Y., Han, S.W., Chen, Y., Zhang, Z.Z., Li, H.W., Li, W.Q., et al., 2021. Characteristics and source apportionment of volatile organic compounds (VOCs) at a coastal site in Hong Kong. Sci. Total Environ. 777, 146241 https://doi.org/10.1016/j.scitotenv.2021.146241.

Wang, H.L., Lou, S.R., Huang, C., Qiao, L.P., Tang, T.B., Chen, C.H., et al., 2014. Source profiles of volatile organic compounds from biomass burning in Yangtze River delta, China. Aerosol Air Qual. Res. 14 (3), 818–828. https://doi.org/10.4209/aaqr.2013.05.0174.

Wang, H.M., Zheng, J.H., Yang, T., He, Z.C., Zhang, P., Liu, X.F., et al., 2020. Predicting the emission characteristics of VOCs in a simulated vehicle cabin environment based on small-scale chamber tests: parameter determination and validation. Environ. Int. 142, 105817 https://doi.org/10.1016/j.envint.2020.105817.

Wang, Z.Y., Yu, H.F., Liang, W.Q., Wang, F., Wang, G., Chen, D., et al., 2022a. Ensemble source apportionment of air pollutants and carbon dioxide based on online measurements. J. Clean. Prod. 370, 133468 https://doi.org/10.1016/j.jclepro.2022.133468.

Wang, S.Y., Zhao, Y.L., Han, Y., Li, R., Fu, H.B., Gao, S., et al., 2022b. Spatiotemporal variation, source and secondary transformation potential of volatile organic compounds (VOCs) during the winter days in Shanghai, China. Atmos. Environ. 286, 119203 https://doi.org/10.1016/j.atmosenv.2022.119203.

Wang, T.T., Tao, J., Li, Z., Lu, X., Liu, Y.L., Zhang, X.R., et al., 2024. Characteristic, source apportionment and effect of photochemical loss of ambient VOCs in an emerging megacity of Central China. Atmos. Res. 305, 107429 https://doi.org/10.1016/j.atmosres.2024.107429.

Wang, Y., Yun, Q.Q., Zhu, L.Y., Zhang, L.P., 2022c. Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on Himawari-8 using a self-adaptive geospatially local model. Geosci. Front. 13 (1), 101286 https://doi.org/10.1016/J.GSF.2021.101286.

Wang, S.Y., Ren, Y., Xia, B.S., 2023. $PM_{2.5}$ and O_3 concentration estimation based on interpretable machine learning. Atmos. Pollut. Res. 14 (9), 101866 https://doi.org/10.1016/j.apr.2023.101866.

Wu, Y.T., Liu, B.S., Meng, H., Dai, Q.L., Shi, L.Y., Song, S.J., et al., 2023a. Changes in source apportioned VOCs during high O_3 periods using initial VOC-concentration-dispersion normalized PMF. Sci. Total Environ. 896, 165182 https://doi.org/10.1016/j.scitotenv.2023.165182.

Wu, C., Ju, Y.C., Yang, S., Zhang, Z.W., Chen, Y.X., 2023b. Reconstructing annual XCO_2 at a 1 km× 1 km spatial resolution across China from 2012 to 2019 based on a spatial CatBoost method. Environ. Res. 236, 116866 https://doi.org/10.1016/j.envres.2023.116866.

Xu, H., Ge, Y., Zhang, C., Wang, Z.Y., Xu, B., Zhao, H., et al., 2023. Machine learning reveals the effects of drivers on $PM_{2.5}$ and CO_2 based on ensemble source apportionment method. Atmos. Res. 295, 107019 https://doi.org/10.1016/j.atmosres.2023.107019.

Yu, S.J., Wang, S.B., Xu, R.X., Zhang, D., Zhang, M., Su, F.C., et al., 2022. Measurement report: intra- and interannual variability and source apportionment of volatile organic compounds during 2018–2020 in Zhengzhou, central China. Atmos. Chem. Phys. 22 (22), 14859–14878. https://doi.org/10.5194/acp-22-14859-2022.

Yuan, B., Shao, M., Lu, S.H., Wang, B., 2010. Source profiles of volatile organic compounds associated with solvent use in Beijing, China. Atmos. Environ. 44, 1919–1926. https://doi.org/10.1016/j.atmosenv.2010.02.014.

Zhang, Z.C., Xu, B., Xu, W.M., Wang, F., Gao, J., Li, Y., et al., 2022. Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on $PM_{2.5}$ pollution. Environ. Res. 212, 113322 https://doi.org/10.1016/j.envres.2022.113322.

Zhang, X., Fan, M.T., Shao, S., Song, X.Q., Wang, H., 2023. Socioeconomic drivers and mitigating strategies of volatile organic compounds emissions in China's industrial sector. Environ. Impact Asses 101, 107102. https://doi.org/10.1016/j.eiar.2023.107102.

Zhao, D.D., Liu, G.J., Xin, J.Y., Quan, J.N., 2020. Haze pollution under a high atmospheric oxidization capacity in summer in Beijing:insights into formation mechanism of atmospheric physicochemical processes. Atmos. Chem. Phys. 20 (8), 4575–4592. https://doi.org/10.5194/acp-2019-966.

Zheng, H., Kong, S.F., Chen, N., Niu, Z.Z., Zhang, Y., Jiang, S.N., et al., 2021. Source apportionment of volatile organic compounds: implications to reactivity, ozone formation, and secondary organic aerosol potential. Atmos. Res. 249, 105344 https://doi.org/10.1016/j.atmosres.2020.105344.