Multi-Machine Learning Approaches to Modeling Small-Scale Source Attribution of Ozone Formation

Zheng Xiao[1,2], Yifeng Lu[3], Guangli Xiu[1,2]

[1]State Environmental Protection Key Lab of Environmental Risk Assessment and Control on Chemical Processes, School of Resources & Environmental Engineering, East China University of Science and Technology, Shanghai 200237, PR China
[2]Shanghai Environmental Protection Key Laboratory for environmental standard and risk management of chemical pollutants, School of Resources & Environmental Engineering, East China University of Science and Technology, Shanghai 200237, PR China
[3]Shanghai Chemical Industry Park Administration Committee, Shanghai 201507, P.R. China

Correspondence to: Guangli Xiu (xiugl@ecust.edu.cn)

Abstract. Accurate source apportionment of ozone (O_3) precursors is crucial for implementing scientific O_3 control strategies. While traditional approaches rely on complex calculations of volatile organic compounds (VOCs) and meteorological parameters, their applicability in real-time scenarios remains limited. Taking the Shanghai chemical industrial park as an example, we propose a novel two-step machine learning (ML) approach that integrates positive matrix factorization (PMF) with other ML methods to systematically quantify the spatiotemporal impacts of VOCs on O_3 formation. Analysis of high-frequency data from 12 VOC monitoring stations (2021-2023) using six ML models revealed XGBoost as the optimal predictor ($R2=0.644$) for local VOC emissions. By combining SHapley Additive exPlanations (SHAP) with ML modeling, we precisely evaluated VOC-O_3 relationships and located emission sources. Results identified solvent use (SU) and fuel evaporation (FE) as primary O_3 formation contributors, followed by combustion sources (CS) and vehicle emissions (VE). PMF analysis further distinguished six VOC sources: petrochemical processes (PP), FE, CS, SU, polymer fabrication (PF), and VE. Temporal analysis revealed seasonal variations, with CS and FE dominant in spring/summer, while PF prevailed in autumn. This innovative framework demonstrates exceptional capability for rapid source identification and precise contribution quantification, establishing a new paradigm for high-resolution O_3 source apportionment.

1 Introduction

Ozone (O_3) pollution has become a significant environmental issue, posing serious threats to human health and ecosystems worldwide (Long et al., 2023; Sharma et al., 2023; Masui et al., 2023; Sharma et al., 2024). In particular, industrial parks, which are characterized by high levels of anthropogenic emissions, serve as critical hotspots for the formation of ground-level O_3 due to the abundance of precursor pollutants such as volatile organic compounds (VOCs) and nitrogen oxides (NO_x) (Pinthong et al., 2022; Kim et al., 2023; He et al., 2024). From the current point of view, the pollution characteristics and sources of VOCs in industrial parks are relatively more complex (Cao et al., 2024). Understanding the contribution of VOCs

to O_3 formation is essential for developing effective mitigation strategies aimed at reducing air pollution and improving air quality.

At present, VOCs source identification technology mainly includes source emission inventory (Lu et al., 2025), chemical transport models (CTMs) (Choi et al., 2020; Wang et al., 2023) and receptor models (Wu et al., 2023). Traditionally, the quantification of VOC contributions to O_3 formation has relied heavily on CTMs, which require detailed knowledge of atmospheric chemistry and complex computational resources (Li et al., 2014). However, these models often suffer from uncertainties related to emission inventories and chemical mechanisms (Sharma et al., 2017; Baklanov and Korsholm, 2008). In addition, the detailed input data and computing power requirements of CTMs leave some areas for improvement (Nelson et al., 2023). The receptor model has the advantages of fewer hardware requirements, higher precision, and easy configuration, among which the positive matrix factorization (PMF) model has been widely used to identify VOC emission sources (Tan et al., 2021; Yang et al., 2023). PMF models are typically combined with the O_3 formation potential (OFP) calculated by the maximum incremental reactivity (MIR) value of VOC species (Carter, 2010) to assess the relationship between VOCs and O_3 concentrations (Xiao et al., 2024). As the key parameter MIR value used in this method is usually calculated based on 39 cities in the United States where O_3 exceeds the standard, whether the MIR value can fully reflect the contribution of VOCs to O_3 under complex atmospheric pollution conditions in China is controversial (Zhang et al., 2021a). Furthermore, conventional methods like PMF, while effective for source categorization, face critical limitations: they cannot differentiate rapid O_3 formation via photostationary state perturbations (e.g., alkene depletion) from slower HOx-mediated pathways (e.g., alkane oxidation) (Sillman, 1999). This mechanistic gap introduces spatiotemporal biases when quantifying source contributions in chemically complex environments like coastal petrochemical zones. To address these challenges, our study integrates interpretable machine learning (ML) with PMF, explicitly resolving fast vs. slow O_3 production pathways while leveraging the study area's unique spatial and industrial characteristics.

In recent years, more ML techniques have emerged as powerful tools for analyzing complex datasets and predicting environmental phenomena (Salcedo-Sanz et al., 2024; Essamlali et al., 2024). However, the "black box" nature of ML models makes their results difficult to interpret and generalize to real scenarios (Guidotti et al., 2018). Recently, the SHapley Additive Interpretation (SHAP) algorithm has been applied to solve these problems (Louhichi et al., 2023; Li et al., 2024). Novel ML approaches can provide robust predictions with less reliance on mechanistic details, making them attractive alternatives or complementary methods to traditional ML models. To address these gaps, we integrate PMF with other advanced ML methods to develop a two-step ML framework. This approach not only enhances the efficiency and accuracy of data analysis but also addresses the limitations of traditional PMF methods under complex environmental conditions, providing a more robust solution for pollution source identification and contribution analysis.

In the past, researchers have used different ML models to study the transformation mechanism of O_3 and its precursors (Cheng et al., 2024; Kuo and Fu, 2023). More recently, ML-coupled receptor models have been used to better identify and quantify the drivers of pollutants. For example, Cheng et al. assessed the impact of emission sources on O_3 formation by combining PMF with four ML models. Chen et al. used PMF model combined with Category Boosting (CatBoost) model and Shapley

65 additive interpretation algorithm to quantify the influence of pollution sources and meteorological factors on VOCs. Ning et al. constructed a cross-stacked ensemble learning model (CSEM) to predict O_3 concentrations under different NOx and VOCs emission reduction scenarios. Our study focuses on a coastal petrochemical industrial zone in Shanghai, where proximity to the ocean and dense aggregation of ethylene crackers and aromatics plants create a unique microenvironment. High humidity and solar irradiance amplify atmospheric oxidative capacity (Zhang et al., 2021), while emissions of short-lived unsaturated

70 VOCs (e.g., propylene, 1,3-butadiene) drive rapid O_3 formation via alkene-NO_x reactions within <50 km under sea-land breezes (3–5 m/s) (Zhang et al., 2025; Zhao et al., 2022; Wang et al., 2018). Additionally, our study pioneers high-resolution source attribution (1 km × 1 km) in a coastal petrochemical cluster in Shanghai, a methodology particularly suited to resolve localized ozone production under elevated regional backgrounds (>80 ppb). By leveraging the unique microenvironment of dense ethylene crackers and aromatics plants adjacent to the ocean, we isolate rapid photochemical processes driven by short-lived,

75 highly reactive VOCs (e.g., propylene, 1,3-butadiene; atmospheric lifetime <6 h) that dominate O_3 formation within <50 km. Considering the advantages of the diverse types and complex sources of VOCs in chemical industrial parks, this characteristic enhances model robustness and improves the model's feature extraction capabilities. This study selected a chemical industrial park in Shanghai as an ideal case for model training. In this study, we present a novel small-scale approach that integrates multiple ML models to quantify the impact of VOCs on O_3 formation with unprecedented spatial resolution. Our methodology

80 harnesses the analytical power of ML algorithms to process high-frequency data from 12 strategically positioned VOC boundary monitoring stations, enabling rapid and accurate source identification at a fine-grained spatial scale previously unattainable through conventional methods. The analytical framework consists of three sophisticated components: First, we conduct a systematic evaluation of diverse ML algorithms to identify the optimal model configuration, ensuring robust predictive performance at the local scale. Subsequently, we leverage the interpretable ML technique SHAP to quantitatively

85 assess VOC-O_3 relationships and precisely pinpoint emission sources with high spatial accuracy. Finally, we develop an innovative hybrid approach that combines PMF for source apportionment with ML-SHAP analysis to achieve rapid and precise identification of key pollution sources contributing to O_3 formation at the facility level. This advanced methodological framework demonstrates significant advantages in both efficiency and spatial precision: it enables swift identification of specific emission sources while maintaining high accuracy in quantifying their individual contributions to O_3 formation. The

90 approach transcends traditional analytical limitations by offering a powerful tool for high-resolution source traceability, thereby providing crucial support for implementing targeted and effective O_3 control strategies at the facility or district level. Moreover, this novel integration of multiple analytical techniques establishes a new paradigm for addressing complex air quality challenges through sophisticated data-driven approaches that bridge the gap between regional-scale analysis and facility-level source identification.

2 Material and methods

2.1 Study area and data details

The Jinshan District is located in the southwestern part of Shanghai, along the northern shore of Hangzhou Bay (HZB) at geographical coordinates 30°40′–30°58′N and 121°–121°25′E, covering a total land area of 586.05 km² (Fig. 1). It is recognized as an important resource-based city. The Jinshan Industrial Zone, situated in the southeastern part of Jinshan District, serves as a fine chemical industrial park. Within and surrounding the industrial zone, 12 monitoring stations have been established to assess the environmental air quality in the area (as detailed in Table S1). The distribution of these monitoring stations is illustrated in Figure 1. Hourly O_3 concentration data utilized in this study were obtained from the Atmospheric Environmental Monitoring Routine Data Management System of the Shanghai Environmental Monitoring Station (https://10.87.87.122:7890/SemcDataExamine). The concentration data for VOCs were sourced from the 12 monitoring stations and can be accessed through the Intelligent Analysis System for VOC Emissions and Pollution Source Tracing in Key Industrial Parks of Shanghai (https://220.196.88.110:8088/GYQ/index.html#/Login). The dataset covers a collection period from January 1, 2021, to December 31, 2023, with measurements recorded on an hourly basis.

Two types of VOC monitoring instruments were deployed at the boundary observation stations. The first type utilized gas chromatography-flame ionization detector (GC-FID) technology, featuring low-carbon (C2-C5) and high-carbon (C6-C12) analyzers (Synspec GC955-615/815, Juguang Technology Co., Ltd.; A11000/A21022, Chromatotec Inc., France; Spectra SYS GC3000-315L/H, Pu Yu Technology Development Co., Ltd.), and enabled automatic hourly measurements of 89 VOC species. The second type employed a combined GC-FID and mass spectrometry (GC-FID/MS) approach, which included FID detection for C2-C5 aromatic hydrocarbons and mass spectrometry detection for other compounds. The instrument operated by sampling at 30 L/min for the initial 10 minutes each hour, utilizing a cryogenic cold trap for sample preservation before separation and detection on specific chromatographic columns. Quality assurance and quality control (QA/QC) protocols adhered to the "Technical Specifications for Operation and Quality Control of Continuous Automatic Monitoring Systems for Gaseous Pollutants in Ambient Air" (HJ 818-2018), issued by China's Ministry of Ecology and Environment. Daily checks ensured data completeness and chromatogram integrity, with any detected abnormalities prompting immediate on-site maintenance. Routine data audits involved the removal of abnormal data, while measurement accuracy was verified biweekly, with calibration curves, method detection limits (MDLs), and instrument precision assessed quarterly. Standard gas accuracy checks showed relative errors below 20%, and in blank tests, absolute errors were less than 0.3 ppbv. Calibration curves were established using five standard gases (1, 5, 10, 15, and 20 ppbv), yielding correlation coefficients greater than 0.995, and MDLs for PAMS and TO-14 species were maintained at or below 0.3 ppbv and 0.5 ppbv, respectively. Throughout the study period, a total of 26,280 sets of VOC data were collected, resulting in the identification of 36 distinct VOC species, comprising 12 alkanes, 7 alkenes, 11 aromatics, and 6 halogenated hydrocarbons after data screening.

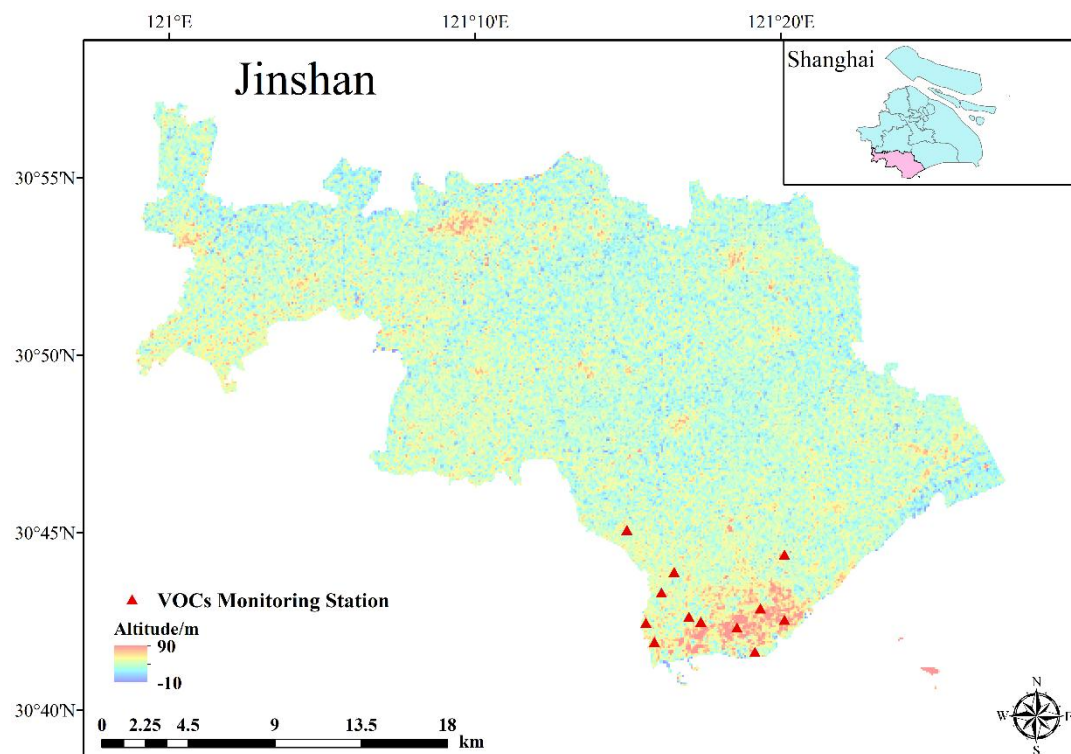



Fig. 1. Study area and sampling sites. The triangles indicate the 12 sites utilized for the evaluation of the model.

2.2 Positive matrix factorization model (PMF)

130 The PMF 5.0 software, developed by the United States Environmental Protection Agency (EPA), is widely utilized to assess and quantify the contributions from various sources to samples based on their chemical composition or unique fingerprints. Initially introduced by Paatero in 1997 at the University of Helsinki, this model decomposes the sample matrix (which is non-negative) into two distinct matrices: the source contribution matrix (g) and the source component spectral matrix (f). The least squares method is subsequently employed to estimate the contribution rates and identify major pollution sources, with the

135 objective of minimizing the discrepancy between the calculated Q-value and the theoretical Q-value. The formulation of the PMF model is represented by Eq. (1):

$$e_{ij} = x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj} \tag{1}$$

where X_{ij} represents the concentration of jth species in the ith sample, g_{ik} represents the concentration of kth source in the ith sample, f_{kj} represents the mass percentage of jth species in the kth sample, and e_{ij} represents the residue factor for j^{th} species in

140 the ith sample.

In the PMF model, factor contributions and source fingerprint spectra are derived by minimizing the objective function (Q),

as described in Eq. (2):

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} ((x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj})/u_{ij})^2 \tag{2}$$

The uncertainty (u_{ij}) of the jth species in the ith sample is calculated by considering the minimum detection limit (MDL) and

145 error scores of each species.

To determine the PMF, equation-based uncertainty is employed in the calculations. Specifically, when the concentration is

equal to or below the method detection limit (MDL), the uncertainty (Unc) is calculated using Eq. (3):

$$Unc = 5/6 \times MDL \tag{3}$$

If the concentration is greater than MDL, it is calculated by Eq. (4):

150 $$Unc = \sqrt{(EF \times concentration)^2 + (0.5 \times MDL)^2} \tag{4}$$

The error score (EF) represents the percentage of uncertainty variables.

2.3 Machine learning models

To explore data characteristics and select the optimal ML model, we utilized data from the Jinshan Industrial Zone covering

the period from January 2021 to December 2023 for training and evaluating ML models. In this study, hourly data of O_3 and

155 total volatile organic compounds (TVOCs) concentrations from 12 monitoring stations were input into the ML models.

Subsequently, the ML models were integrated with the SHAP (SHapley Additive exPlanations) (Section S1.2 for model

details) framework to obtain SHAP values for the 12 stations, quantifying their contributions to O_3 formation. The stations

contributing the most to O_3 were identified, and the source of characteristic VOC data was recognized using the PMF model.

The emission factors derived from the PMF analysis, along with hourly O_3 concentration data, were then input into the ML

160 models for SHAP analysis. Finally, the optimal ML model and the most severe pollution sources were identified.

In this study, we implemented six ML models, including Decision Tree Regression (DTR), Random Forest Regression (RF),

Support Vector Regression (SVR), XGBoost Model, CatBoost Model, and LightGBM Model (refer to Section S1.1 for model

details). Compared to complex deep learning models, these six ML models are more adept at capturing variations in internal

parameters and providing variable interpretations (Pichler and Hartig, 2023; Kaur et al., 2020). Utilizing a large-scale dataset

165 from this research, Bayesian Optimization (Robin et al., 2021) was employed to determine the optimal values of a range of

hyperparameters for the ML models. Ultimately, through iterative adjustments in training strategies and multiple random experiments, we selected the model with the highest R² from the six ML models, along with its corresponding parameter combinations, as the optimal ML model. Additionally, to assess the robustness and stability of the models, we employed Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as evaluation metrics.

170 R² is an indicator that measures the overall goodness of fit of a regression model, and its calculation is shown in Eq. (5):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i represents the actual observed values, $\hat{y}_i$ denotes the predicted values, $\bar{y}$ is the mean of the actual observed values, and n indicates the total number of observations.

MAE is the average of the absolute differences between the predicted values and the actual values, as calculated in Eq. (6):

175 $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

where y_i is the actual observed value, $\hat{y}_i$ is the predicted value, and n is the total number of observations.

RMSE is the square root of the average of the squared differences between the predicted values and the actual values, as shown in Eq. (7):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

180 where y_i represents the actual observed values, $\hat{y}_i$ denotes the predicted values, and n is the total number of observations. This metric provides insight into the model's accuracy by giving a higher weight to larger errors, making it sensitive to outliers.

To ensure data integrity, a thorough check for missing values was conducted on the original observational data prior to its input into the machine learning models. Subsequently, these missing values were imputed using the RF model, and validation was performed using the available observational data. The resulting dataset was then transformed into a supervised learning

185 dataset with time-dependent features. To guarantee the accuracy of the experiments, 70% of the dataset was consistently used as the training set and 30% as the test set for each ML model. Hyperparameter tuning was performed utilizing the Bayesian Optimization technique. Furthermore, to mitigate the influence of model randomness on the test set results, all ML models underwent 10-fold cross-validation experiments.

In this study, the ML models and the SHAP algorithm were primarily implemented using Python 3.6 and Anaconda 4.5

190 platforms. The research methodology framework used in this study is shown in Fig. 2.

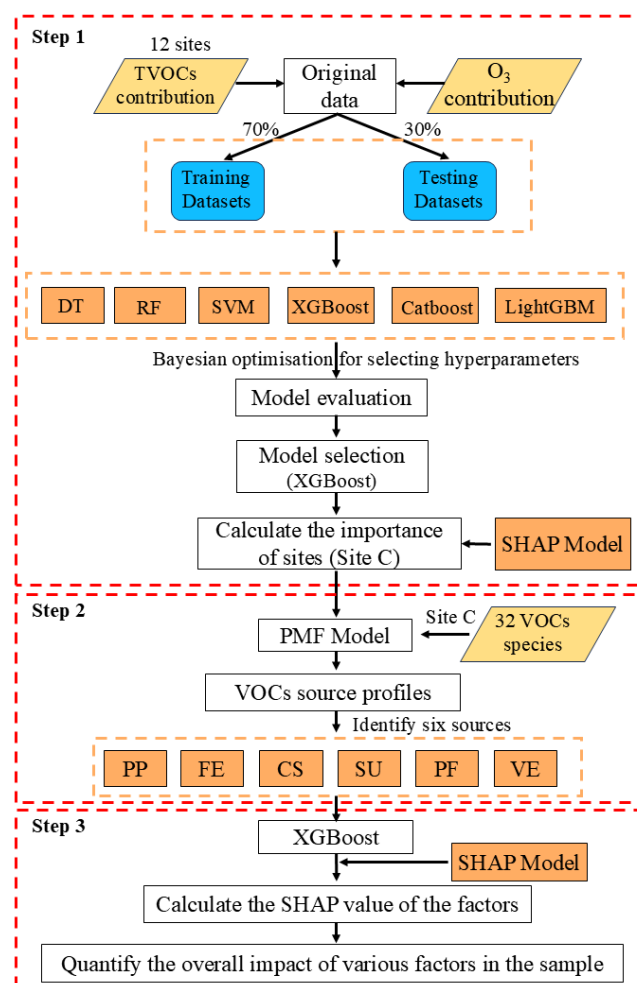



Fig. 2. Workflow of the development of the model. The framework comprises three sequential phases: (1) Model Training and Optimization: Multiple machine learning models including Decision Tree (DT), Random Forest (RF) Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Category Boosting (CatBoost), and Light Gradient Boosting Machine (LightGBM) were trained using TVOCs and O₃ contribution data from 12 monitoring stations (70% training set/30% testing set). Bayesian hyperparameter optimization and rigorous model evaluation were implemented to identify the optimal XGBoost model. Subsequently, SHapley Additive exPlanations (SHAP) analysis was employed to pinpoint critical monitoring stations (e.g., Site C). (2) VOCs Source Apportionment: Positive Matrix Factorization (PMF) was applied to 32 VOCs species at key stations, resolving six emission sources: petrochemical processes (PP), fuel evaporation (FE), combustion sources (CS), solvent use (SU), polymer fabrication (PF), and vehicle emissions (VE). (3) Factor Impact Quantification: SHAP values derived from the XGBoost model were calculated to dynamically quantify the contributions of source categories, and chemical species to O₃ formation. This systematic workflow integrates a cohesive "data modeling-source identification-attribution analysis" chain, demonstrating strong applicability for high-resolution pollution source tracing studies.

3 Results and discussions

3.1 Spatiotemporal characterization of ozone and VOCs Concentrations: a multi-scale analysis for enhanced ML performance

A comprehensive understanding of the spatiotemporal characteristics of the input dataset is fundamental for optimizing machine learning model interpretability and performance. Our analysis of VOC and O_3 concentrations across multiple monitoring sites from 2021 to 2023 reveals distinct patterns that inform our ML-based source attribution approach. The temporal evolution of O_3 concentrations shows a notable progression, with mean values of 64.45 μg/m³, 68.79 μg/m³, and 68.90 μg/m³ recorded in 2021, 2022, and 2023, respectively (Fig. 3a). The relatively lower O_3 levels observed in 2021 coincide with the COVID-19 pandemic period, when reduced anthropogenic activities, particularly decreased vehicular traffic and industrial operations, significantly altered urban emission patterns in Shanghai (Lu et al., 2023). A marked reduction in VOC concentrations was observed across all monitoring sites in 2022, primarily attributed to the implementation of stringent emission control measures by local regulatory authorities (Xiao et al., 2024). The seasonal analysis reveals distinctive O_3 formation patterns in the study area (Fig. 3b). O_3 concentrations exhibit a rapid acceleration during spring, reaching maximum levels in May, followed by a notable decline during the summer months (June-August). This pattern differs from typical urban environments, as Shanghai's unique high-pressure meteorological conditions drive peak daytime O_3 levels in late May with unprecedented rates of increase (Chang et al., 2021). Notably, elevated VOC concentrations coincide with the June industrial maintenance period, providing critical temporal markers for our ML-based source attribution.

The diurnal profile analysis (Fig. 3c) reveals sophisticated photochemical patterns that enhance our ML model's temporal resolution. O_3 concentrations follow a characteristic curve, initiating a rapid increase at 08:00, sustaining growth until reaching peak levels at 15:00, followed by a gradual decline. This pattern aligns with the intensification of photochemical reactions driven by increasing solar radiation. The observed dual-peak pattern in VOC concentrations (06:00-08:00 and 19:00-22:00) corresponds to industrial operational cycles, providing essential precursor availability for daytime O_3 formation. This rich temporal diversity in our dataset significantly enhances the ML model's capability to capture complex source-receptor relationships.

The correlation analysis provides crucial insights for optimizing our ML framework's feature selection and interpretation capabilities (Fig. 3d). The heterogeneous correlations between VOCs and O_3 across monitoring sites reveal complex source-receptor relationships: while most sites exhibit negative correlations, Sites D and K show positive associations. Notably, Sites C (r = -0.12) and L (r = -0.081) demonstrate the strongest negative correlations, marking them as critical locations for detailed feature importance analysis in our ML framework. These diverse correlation patterns enhance our model's ability to capture local-scale emission-concentration relationships, crucial for precise source attribution.

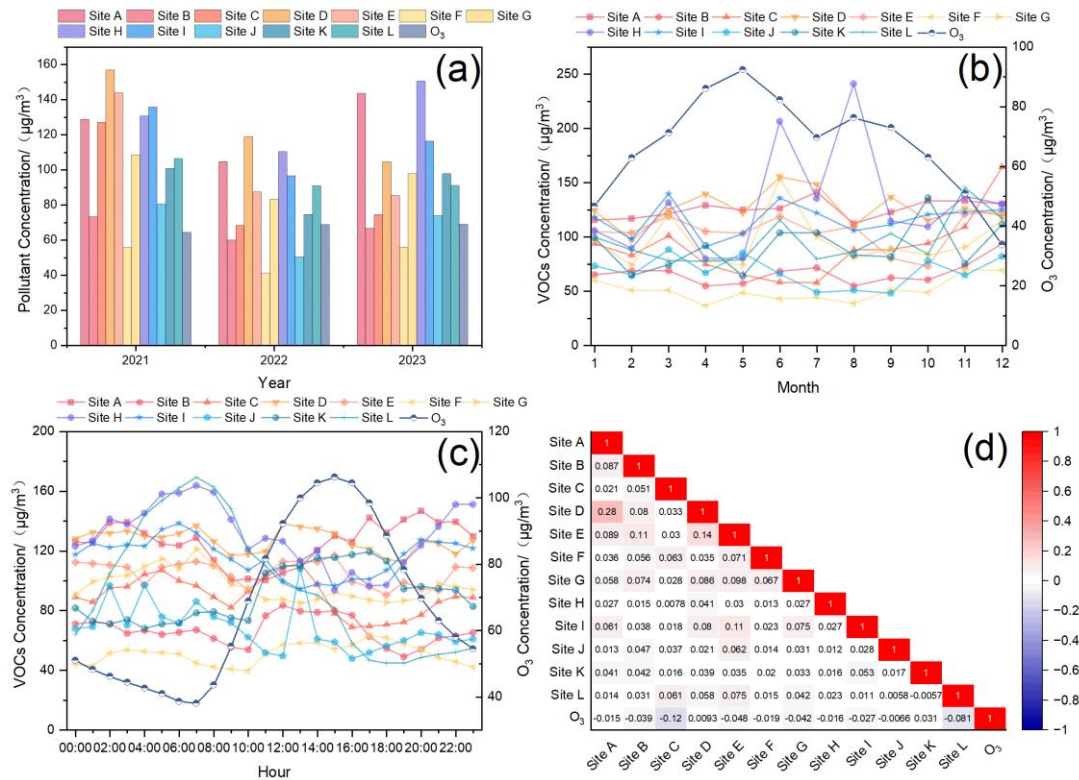



Fig. 3 Temporal and spatial dynamics of VOCs and O₃, along with their correlations, as observed at 12 monitoring stations. (a) Mean concentration variations of O₃ and VOCs from 2021 to 2023. (b) Monthly profiles of O₃ and VOC concentrations. (c) Daily profiles of O₃ and VOC concentrations. (d) The correlation matrix between O₃ and VOCs measurements at the 12 monitoring sites.

3.2 Comparison of ML models

3.2.1 Model performance evaluation and selection of optimal ML algorithm

The selection and evaluation of appropriate machine learning algorithms are fundamental to ensuring robust and reliable analytical outcomes, particularly when dealing with complex environmental datasets (Liu et al., 2022). In this investigation, we implemented and systematically evaluated six state-of-the-art machine learning algorithms: DT, RF, SVM, XGBoost, CatBoost, and LightGBM. These models were rigorously trained and tested using spatially distributed VOC and O₃ monitoring data from multiple sampling sites.

The comparative performance metrics of these models are comprehensively presented in Table 1. Notably, R^2 values obtained through ten-fold cross-validation demonstrated remarkable consistency with those derived from the independent test dataset. This concordance strongly indicates the robust generalization capabilities of our machine learning framework, suggesting effective pattern recognition within the data while avoiding overfitting issues. Among the evaluated algorithms, the XGBoost model demonstrated superior predictive performance across all assessment metrics. Specifically, it achieved the lowest MAE

of 13.828, Mean Absolute Percentage Error (MAPE) of 0.445, and RMSE of 11.620, coupled with the highest R² value of 0.637.

Further validation through scatter plot analysis (Fig. S1) revealed that the XGBoost model exhibited significantly smaller deviations between predicted and observed values compared to other algorithms, confirming its enhanced predictive accuracy.

255   Based on these comprehensive evaluation results, the XGBoost model was selected as the optimal algorithm for O₃ concentration prediction. Consequently, the subsequent SHAP analysis was conducted exclusively on the XGBoost model to ensure the highest level of interpretability and reliability in source attribution results.

Table 1 Comparison of six ML models based on different evaluation metrics.

Model	10-fold Cross-validation (R²)	R²	MAE	MAPE	RMSE
DT	0.167	0.187	23.142	0.589	29.980
RF	0.345	0.383	21.579	2.918	28.155
SVM	0.330	0.328	25.359	0.618	34.309
XGBoost	0.637	0.644	13.828	0.445	11.620
CatBoost	0.689	0.630	15.601	0.495	17.216
LightGBM	0.675	0.624	14.603	0.493	17.352

(MAE: Mean Absolute Error, MAPE: Mean Absolute Percentage Error, RMSE: Root Mean Squared Error, DT: Decision Tree,
260   **RF: Random Forest, SVM: Support Vector Machine, XGBoost: Extreme Gradient Boosting, CatBoost: Category Boosting,**
LightGBM: Light Gradient Boosting Machine)

3.2.2 Spatial distribution analysis of TVOC contributions using SHAP interpretation

To quantitatively assess the spatial heterogeneity of TVOC contributions to O₃ formation, we conducted a comprehensive SHAP analysis across all monitoring sites within the established CatBoost regression model (Fig. 4). The analysis revealed
265   distinct patterns of TVOC influence, with Site C demonstrating notably dispersed sample distribution patterns, indicating its predominant influence on O₃ formation dynamics. The analysis identified a significant negative correlation between TVOC concentrations at Site C and model predictions, evidenced by the high-magnitude SHAP values concentrated in the negative region. Conversely, Site K exhibited a strong positive correlation, characterized by elevated SHAP values consistently distributed in the positive region, corroborating the correlation analysis findings presented in Section 3.1.

270   The pronounced influence of Site C can be attributed to its strategic location within the southeastern sector of the Jinshan Chemical Zone, where it is surrounded by diverse industrial activities. The site's immediate vicinity encompasses plastic manufacturing facilities to the east, a petroleum products transportation port to the south, chemical production facilities to the west, and a public transportation hub featuring five gas stations to the north. This complex industrial landscape surrounding Site C facilitates intensive O₃ formation while presenting significant challenges for precise source attribution of TVOC
275   emissions.

To validate the robustness of these findings, we extended the SHAP analysis to encompass five additional ML models (Fig. S2). The results consistently identified Site C as the most influential monitoring location across all model analyses, substantiating its critical role in local O₃ formation processes. This convergence of results across multiple ML platforms reinforces the reliability of our spatial analysis and highlights the importance of Site C in understanding regional O₃ pollution

280   patterns.

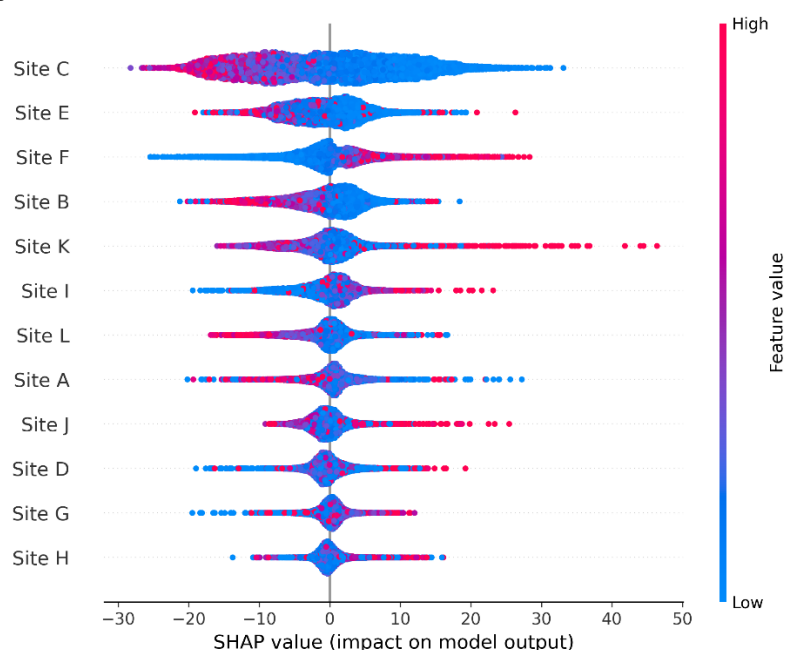



Fig. 4 Feature importance of total volatile organic compounds (TVOC) drivers obtained by XGBoost model. Blue points indicate negative contributions to the prediction, while red points represent positive contributions.

3.3 Source apportionment and temporal-sspatial characterization of VOCs based on the PMF model

285   **3.3.1 Analysis process and results of the PMF model**

The application of the PMF model for the analysis of VOCs at Site C offers valuable insights into source apportionment and pollution characteristics. The species selection process adhered to specific criteria (Liu et al., 2016; Hui et al., 2019): (1) species with missing sample rates exceeding 25% or concentrations below 35% of the method detection limits (MDLs) were excluded; (2) highly reactive compounds were omitted unless serving as specific tracers for particular sources. Ultimately, 36

290   VOC species were selected for model input. Based on their signal-to-noise ratio (S/N) and model fit, species were classified as "strong," "weak," or "bad" in terms of their computational significance. Specifically, species with $S/N \leq 0.2$ were labeled as "bad," those with $0.2 < S/N \leq 0.6$ or poor fit as "weak," and $S/N > 0.6$ as "strong." In total, 30 species were deemed "strong," 2 "weak," and 4 "bad." The "bad" species were excluded from model calculations due to their concentration uncertainty, while the uncertainty for "weak" species was tripled to reduce their influence. Ultimately, 32 pollutants were included in the model.

295 To ascertain the optimal number of sources, 20 iterations with varying random seeds were conducted to evaluate the stability of the Q value across solutions ranging from 3 to 10 factors. A 6-factor solution was selected as the rate of Q value reduction significantly diminished beyond this point, and the model results remained interpretable. The Q (true)/Q (robust) value of 1.23 fell within the acceptable range of <1.5 (Hui et al., 2020), while the Q (robust)/Q (theoretical) value of 1.1 indicated proximity to 1. The standardized residuals for each factor ranged from -3 to 3, demonstrating a satisfactory fit between the model

300 predictions and observed data. The six sources identified using the PMF model are as follows: petrochemical process (PP), fuel evaporation (FE), combustion source (CS), solvent use (SU), polymer fabrication (PF), and vehicle emission (VE). The source profiles are illustrated in Fig. 5.

Factor 1 exhibited the highest concentration of propylene (60.60%), suggesting its predominant source is local petrochemical processes. This finding aligns with previous studies that link propylene emissions to petrochemical activities in industrial zones

305 (Washenfelder et al., 2010; Ragothaman and Anderson, 2017). Therefore, Factor 1 was determined to be PP. Factor 2 was characterized mainly by C2-C5 alkanes (47%) and toluene (13.68%); they are typical components of gasoline and diesel (Mu et al., 2023). Thus, Factor 2 is defined as a FE.

Factor 3 had a high percentage of propane (69.36%), ethene (12.17%), and ethane (5.12%), which conformed to the emission characteristics of CS (Song et al., 2021; Chen et al., 2024). The relative contribution rate of aromatic hydrocarbons in Factor

310 4 reached 69.12%. Currently, aromatic hydrocarbons are widely used as solvents in industrial production (Zhang et al., 2021b; Mukhamatdinov et al., 2020). Factor 4 was defined as SU.

Factor 5 was primarily characterized by ethene (92.59%), with its concentration far exceeding that of other species. It indicated that its source was closely related to the polymer manufacturing processes commonly found in nearby production facilities. Previous studies had identified ethene as a major pollutant associated with such industrial activities (Burdett and Eisinger,

315 2017). Therefore, Factor 5 was identified as originating from PF. Factor 6 was characterized by relatively high proportions of C4-C6 alkanes (37.63%) and 1,2-dichloroethane (12.01%), which are important indicators of VE (Song et al., 2021; Chen et al., 2024).

The PMF analysis revealed distinct temporal patterns in source contributions to VOC emissions throughout the study period (Fig. 6). On an annual basis, CS and FE emerged as the primary contributors, accounting for 16.94% and 16.89% of total VOC

320 emissions, respectively. The remaining sources exhibited comparable contributions: SU (16.56%), PF (16.54%), VE (16.54%), and PP (16.52%), indicating a relatively balanced distribution of emission sources in the industrial park.

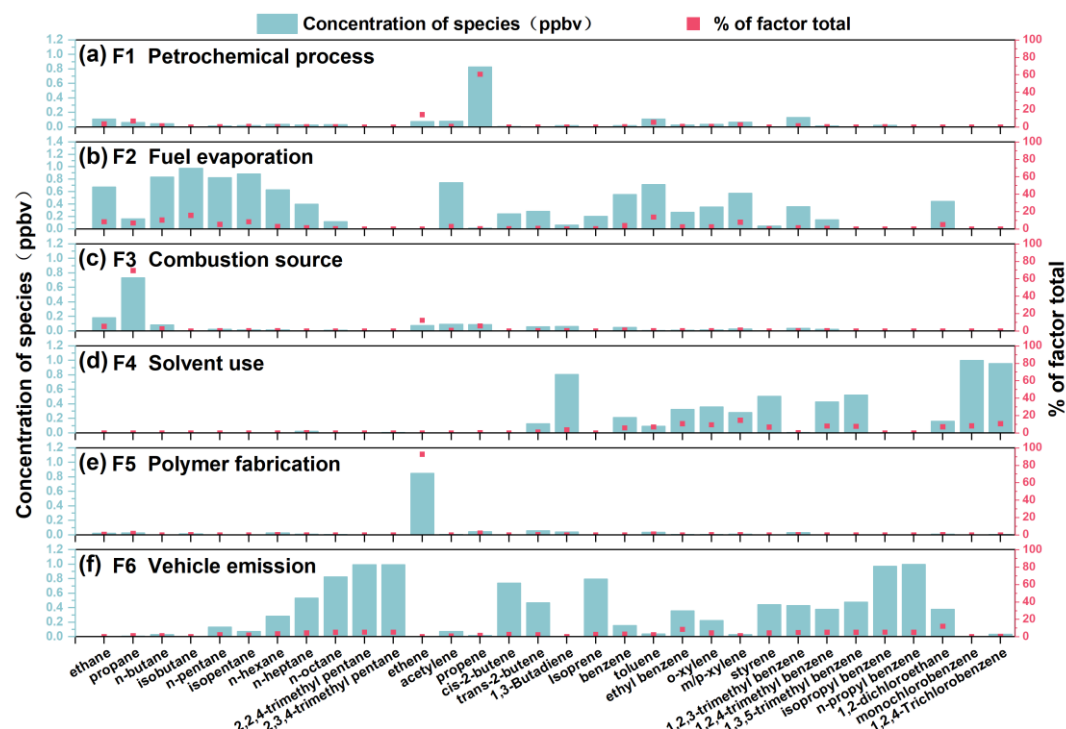



Fig. 5 Source profiles calculated using the PMF model in site C.

3.3.2 Seasonal variation and contributing factors of VOC source distribution

325 Seasonal analysis unveiled significant temporal variations in source contributions. During spring, CS dominated the VOC emissions with a 22.39% contribution, followed by PP (19.59%), PF (16.07%), and VE (14.76%), while FE showed a relatively lower contribution of 12.83%. This pattern shifted notably in summer, where FE became the predominant source (21.68%), accompanied by substantial contributions from SU (18.21%) and VE (17.59%). The autumn period witnessed PF emerging as the primary contributor (22.47%), with CS maintaining a significant presence (16.95%). Winter emissions were primarily

330 attributed to VE (18.64%), SU (18.61%), and FE (18.02%).

The observed seasonal variations align with regional industrial activities and meteorological conditions. The pronounced CS contributions during spring coincide with increased biomass burning activities across industrial parks in China, corroborating findings from previous studies (Yang et al., 2023; Yao et al., 2021; Chen et al., 2022). The elevated FE contributions in summer can be attributed to enhanced fuel volatilization under high temperatures characteristic of the Yangtze River Delta region,

335 subsequently promoting O_3 formation (Xu et al., 2023). The autumn dominance of PF emissions reflects the operational patterns of polymer manufacturing facilities, while the significant winter contribution from VE aligns with previously documented patterns of vehicular emissions in Shanghai's urban areas (Liu et al., 2021; Wang et al., 2022).

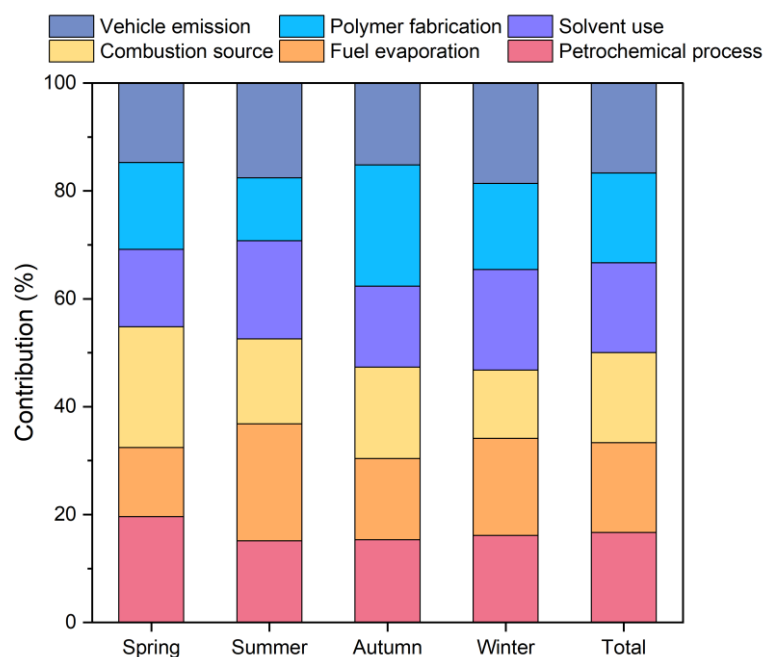



Fig. 6 Influence of various sources on atmospheric VOCs across different seasons.

340 **3.4 Quantitative assessment of source-specific contributions to O$_3$ formation**

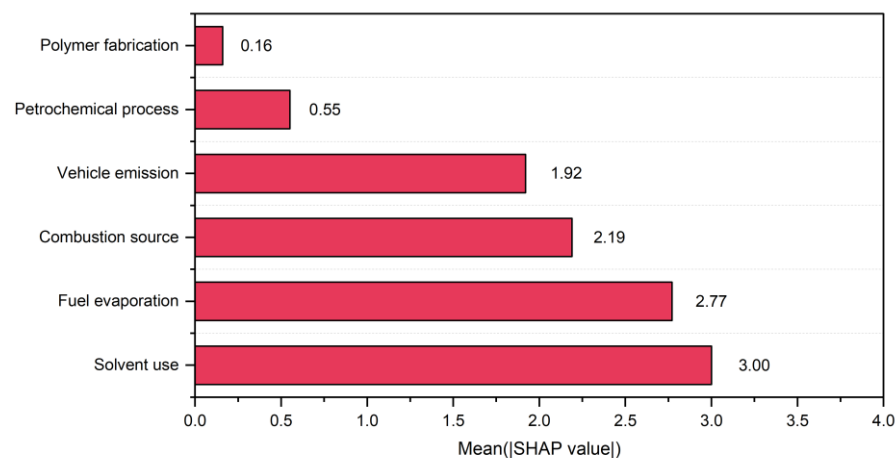
The integrated PMF-SHAP framework was employed to quantitatively evaluate the source-specific contributions to O$_3$ formation in the Jinshan industrial complex during 2021-2023. Through the coupling of XGBoost-SHAP modeling with corresponding SHAP values, we systematically assessed the relative importance of various emission sources in driving O$_3$ pollution dynamics (Fig. 7). The analysis revealed that SU and FE were the predominant contributors to O$_3$ formation,

345 exhibiting SHAP values of 3.00 and 2.77, respectively. This finding underscores the critical role of solvent usage and fuel-related emissions in industrial processes as primary drivers of O$_3$ generation. CS demonstrated moderate influence with a SHAP value of 2.19, followed by VE (1.92) and PP (0.55), while PF exhibited the lowest impact (0.16). These results emphasize the particular significance of SU and FE in O$_3$ pollution control strategies within industrial contexts.

The seasonal decomposition of source contributions through SHAP analysis demonstrated strong concordance with PMF-

350 derived temporal patterns (Fig. 8). CS emerged as the dominant contributor during spring (30.51%), while summer was characterized by substantial contributions from FE (27.25%) and SU (24.48%). The autumn period was dominated by PF emissions (27.83%), whereas SU showed peak influence during winter (31.23%). This temporal heterogeneity in source contributions, particularly the significant TVOC emissions associated with Site C, provides crucial insights for targeted O$_3$ management strategies.

355 These findings have significant implications for policy development and implementation. The results suggest that regulatory attention should prioritize emission controls around Site C, with particular emphasis on seasonal variation in source

contributions. Implementation of season-specific control strategies, especially targeting predominant sources during high-O_3 periods, would optimize the effectiveness of O_3 pollution mitigation efforts in industrial areas.

360   **Fig. 7 Mean SHAP values for six pollution sources affecting ozone levels.**

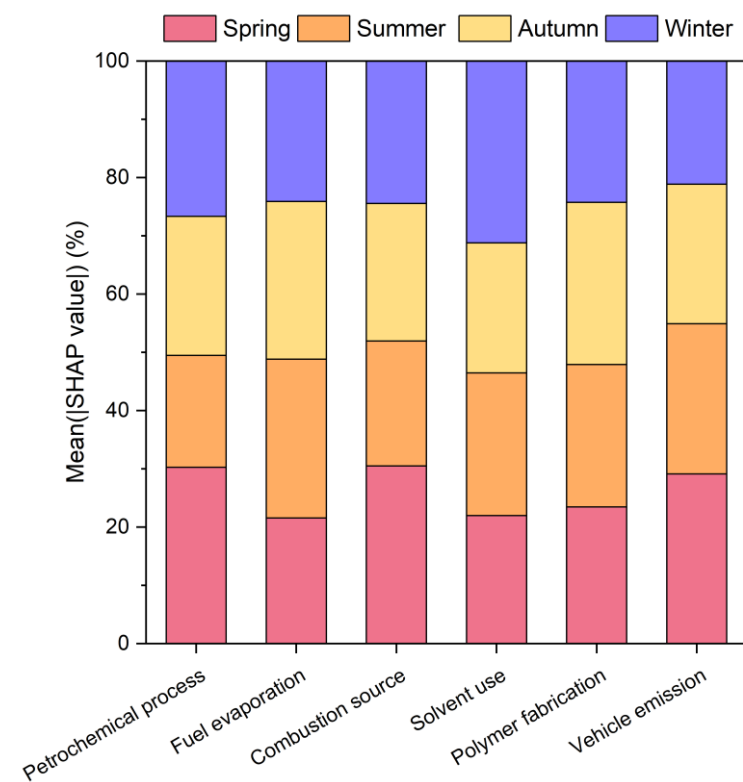



Fig. 8 Proportional contributions of six pollution sources to ozone levels, represented by SHAP values across different seasons.

4. Conclusions

This study presents a novel methodological framework for quantifying VOC contributions to O₃ formation in industrial park environments, integrating advanced ML techniques with traditional source apportionment methods. Through the synergistic combination of ML algorithms, SHAP interpretation, and PMF modeling, we have developed a robust analytical approach that provides unprecedented spatial and temporal resolution in source identification and contribution assessment.

The investigation revealed distinct patterns in VOC-O₃ relationships, with solvent utilization and fuel evaporation emerging as primary drivers of O₃ formation in the industrial complex. The XGBoost model demonstrated superior predictive performance ($R^2 = 0.644$) among the evaluated ML algorithms, while SHAP analysis enabled precise quantification of source-specific contributions. The PMF analysis further delineated six distinct emission sources, exhibiting pronounced seasonal variations in their relative contributions to O₃ formation.

Notably, combustion sources dominated spring emissions (30.51%), while fuel evaporation (27.25%) and solvent use (24.48%) were predominant during summer months. This temporal heterogeneity in source contributions underscores the necessity for season-specific control strategies tailored to industrial operational patterns and meteorological conditions. The significant influence of Site C, characterized by diverse industrial activities, highlights the importance of targeted emission controls in areas with complex source profiles.

These findings provide crucial insights for evidence-based policy development in industrial air quality management. The methodology established herein offers a powerful tool for rapid source identification and precise contribution quantification, enabling the implementation of targeted control strategies at facility-level resolution. Future research directions should focus on expanding the temporal and spatial coverage of monitoring networks, and exploring the application of this framework across diverse industrial settings to enhance its generalizability and predictive capabilities.

5. Appendix A: Acronym glossary

Abbreviation	Full Name
CatBoost	Category Boosting
CS	combustion sources
CSEM	cross-stacked ensemble learning model
CTMs	chemical transport models
DTR	Decision Tree Regression
EF	error score
FE	fuel evaporation
GC-FID	gas chromatography-flame ionization detector
GC-FID/MS	gas chromatography-flame ionization detector and mass spectrometry

LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
MDL	minimum detection limit
MDLs	method detection limits
MIR	maximum incremental reactivity
ML	machine learning
NO_x	nitrogen oxides
O_3	ozone
OFP	ozone formation potential
PF	polymer fabrication
PMF	positive matrix factorization
PP	petrochemical processes
RF	Random Forest
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SU	solvent use
SVM	Support Vector Machine
SVR	Support Vector Regression
TVOCs	total volatile organic compounds
VE	vehicle emissions
VOCs	volatile organic compounds
XGBoost	Extreme Gradient Boosting

385 **Code availability**

Code related to this paper may be requested from the authors.

Data availability

The observational data from Jinshan District, Shanghai, from 2021 to 2023 is confidential. Hourly O_3 concentration data
utilized in this study were obtained from the Atmospheric Environmental Monitoring Routine Data Management System of
390 the Shanghai Environmental Monitoring Station: https://10.87.87.122:7890/SemcDataExamine. The concentration data for
VOCs were sourced from the 12 monitoring stations and can be accessed through the Intelligent Analysis System for VOC

Emissions and Pollution Source Tracing in Key Industrial Parks of Shanghai: https://220.196.88.110:8088/GYQ/index.html#/Login.

Author contributions

ZX, GX, and YL conceived and supervised the study. ZX analyzed the data. ZX wrote the paper with input from GX and YL. GX reviewed and commented on the paper. All authors contributed to discussing the results and revising the draft.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

The authors feel very appreciate for the help of observation from Shanghai environmental monitoring center and Shanghai chemical industry park administration committee.

Financial support

The work is financially supported by grant from National Ministry of Science and Technology (grant nos. 2022YFC3703501 and 2022YFC3703503), Shanghai Jinshan Municipal Bureau of Ecology and Environment (grant nos. Huhuanke-2023-22 and Huhuanke-2023-35).

References

Baklanov, A. and Korsholm, U.: On-line integrated meteorological and chemical transport modelling: advantages and prospectives, Air Pollution Modeling and Its Application XIX, Springer Netherlands, 3-17, https://doi.org/10.1007/978-1-4020-8453-9_1, 2008.

Burdett, I. D. and Eisinger, R. S.: Ethylene polymerization processes and manufacture of polyethylene, Handbook of Industrial Polyethylene and Technology: Definitive Guide to Manufacturing, Properties, Processing, Applications and Markets, 61-103, https://doi.org/10.1002/9781119159797, 2017.

415   Cao, X., Yi, J., Li, Y., Zhao, M., Duan, Y., Zhang, F., and Duan, L.: Characteristics and Source Apportionment of Volatile Organic Compounds in an Industrial Area at the Zhejiang–Shanghai Boundary, China, Atmosphere, 15, 237, https://doi.org/10.3390/atmos15020237, 2024.

Carter, W. P.: Development of the SAPRC-07 chemical mechanism, Atmos. Environ., 44, 5324-5335, https://doi.org/10.1016/j.atmosenv.2010.01.026, 2010.

420   Chang, L., He, F., Tie, X., Xu, J., and Gao, W.: Meteorology driving the highest ozone level occurred during mid-spring to early summer in Shanghai, China, Sci. Total Environ., 785, 147253, https://doi.org/10.1016/j.scitotenv.2021.147253, 2021.

Chen, D., Zhou, L., Wang, C., Liu, H., Qiu, Y., Shi, G., Song, D., Tan, Q., and Yang, F.: Characteristics of ambient volatile organic compounds during spring O3 pollution episode in Chengdu, China, J. Environ. Sci., 114, 115-125, https://doi.org/10.1016/j.jes.2021.08.014, 2022.

425   Chen, W., Xu, X., and Liu, W.: Combined PMF modelling and machine learning to identify sources and meteorological influencers of volatile organic compound pollution in an industrial city in eastern China, Atmos. Environ., 334, 120714, https://doi.org/10.1016/j.atmosenv.2024.120714, 2024.

Cheng, N., Jing, D., Gu, Z., Cai, X., Shi, Z., Li, S., Chen, L., Li, W., and Wang, Q.: Observation-Based Ozone Formation Rules by Gradient Boosting Decision Trees Model in Typical Chemical Industrial Parks, Atmosphere, 15, 600,
430   https://doi.org/10.3390/atmos15050600, 2024.

Cheng, Y., Huang, X.-F., Peng, Y., Tang, M.-X., Zhu, B., Xia, S.-Y., and He, L.-Y.: A novel machine learning method for evaluating the impact of emission sources on ozone formation, Environ. Pollut., 316, 120685, https://doi.org/10.1016/j.envpol.2022.120685, 2023.

Choi, M. S., Qiu, X., Zhang, J., Wang, S., Li, X., Sun, Y., Chen, J., and Ying, Q.: Study of secondary organic aerosol formation
435   from chlorine radical-initiated oxidation of volatile organic compounds in a polluted atmosphere using a 3D chemical transport model, Environ. Sci. Technol., 54, 13409-13418, https://doi.org/10.1021/acs.est.0c02958, 2020.

Essamlali, I., Nhaila, H., and El Khaili, M.: Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review, Sustainability, 16, 976, https://doi.org/10.3390/su16030976, 2024.

440   Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D.: A survey of methods for explaining black box models, ACM. Comput. Surv, 51, 1-42, https://doi.org/10.1145/3236009, 2018.

He, L., Duan, Y., Zhang, Y., Yu, Q., Huo, J., Chen, J., Cui, H., Li, Y., and Ma, W.: Effects of VOC emissions from chemical industrial parks on regional O3-PM2. 5 compound pollution in the Yangtze River Delta, Sci. Total Environ., 906, 167503, https://doi.org/10.1016/j.scitotenv.2023.167503, 2024.

445   Hui, L., Liu, X., Tan, Q., Feng, M., An, J., Qu, Y., Zhang, Y., and Cheng, N.: VOC characteristics, sources and contributions to SOA formation during haze events in Wuhan, Central China, Sci Total Environ, 650, 2624-2639, https://doi.org/10.1016/j.scitotenv.2018.10.029, 2019.

Hui, L., Liu, X., Tan, Q., Feng, M., An, J., Qu, Y., Zhang, Y., Deng, Y., Zhai, R., and Wang, Z.: VOC characteristics, chemical reactivity and sources in urban Wuhan, central China, Atmos. Environ., 224, https://doi.org/10.1016/j.atmosenv.2020.117340, 2020.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J.: Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, Proceedings of the 2020 CHI conference on human factors in computing systems, 1-14, https://doi.org/10.1145/3313831.3376219, 2020.

Kim, S.-J., Lee, H.-Y., Lee, S.-J., and Choi, S.-D.: Passive air sampling of VOCs, O3, NO2, and SO2 in the large industrial city of Ulsan, South Korea: spatial–temporal variations, source identification, and ozone formation potential, Environ. Sci. Pollut. Res., 30, 125478-125491, https://doi.org/10.1007/s11356-023-31109-z, 2023.

Kuo, C.-P. and Fu, J. S.: Ozone response modeling to NOx and VOC emissions: Examining machine learning models, Environ. Int., 176, 107969, https://doi.org/10.1016/j.envint.2023.107969, 2023.

Li, M., Sun, H., Huang, Y., and Chen, H.: Shapley value: from cooperative game to explainable artificial intelligence, Auton. Intell. Syst., 4, 2, https://doi.org/10.1007/s43684-023-00060-8, 2024.

Li, M., Zhang, Q., Streets, D., He, K., Cheng, Y., Emmons, L., Huo, H., Kang, S., Lu, Z., and Shao, M.: Mapping Asian anthropogenic emissions of non-methane volatile organic compounds to multiple chemical mechanisms, Atmos. Chem. Phys., 14, 5617-5638, https://doi.org/10.5194/acp-14-5617-2014, 2014.

Liu, B., Liang, D., Yang, J., Dai, Q., Bi, X., Feng, Y., Yuan, J., Xiao, Z., Zhang, Y., and Xu, H.: Characterization and source apportionment of volatile organic compounds based on 1-year of observational data in Tianjin, China, Environ Pollut, 218, 757-769, https://doi.org/10.1016/j.envpol.2016.07.072, 2016.

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

Liu, Y., Wang, H., Jing, S., Peng, Y., Gao, Y., Yan, R., Wang, Q., Lou, S., Cheng, T., and Huang, C.: 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, 2021.

Long, Y., Wu, Y., Xie, Y., Huang, L., Wang, W., Liu, X., Zhou, Z., Zhang, Y., Hanaoka, T., and Ju, Y.: PM2. 5 and ozone pollution-related health challenges in Japan with regards to climate change, Global Environ. Change, 79, 102640, https://doi.org/10.1016/j.gloenvcha.2023.102640, 2023.

Louhichi, M., Nesmaoui, R., Mbarek, M., and Lazaar, M.: Shapley values for explaining the black box nature of machine learning model clustering, Procedia Comput. Sci., 220, 806-811, https://doi.org/10.1016/j.procs.2023.03.107, 2023.

Lu, B., Zhang, Z., Jiang, J., Meng, X., Liu, C., Herrmann, H., Chen, J., Xue, L., and Li, X.: Unraveling the O3-NOX-VOCs relationships induced by anomalous ozone in industrial regions during COVID-19 in Shanghai, Atmos. Environ., 308, https://doi.org/10.1016/j.atmosenv.2023.119864, 2023.

480 Lu, X., Zhang, D., Wang, L., Wang, S., Zhang, X., Liu, Y., Chen, K., Song, X., Yin, S., and Zhang, R.: Establishment and verification of anthropogenic speciated VOCs emission inventory of Central China, J. Environ. Sci., 149, 406-418, https://doi.org/10.1016/j.jes.2024.01.033, 2025.

Masui, N., Shiojiri, K., Agathokleous, E., Tani, A., and Koike, T.: Elevated O3 threatens biological communications mediated by plant volatiles: A review focusing on the urban environment, Crit. Rev. Environ. Sci. Technol., 53, 1982-2001,
485 https://doi.org/10.1080/10643389.2023.2202105, 2023.

Mu, J., Zhang, Y., Xia, Z., Fan, G., Zhao, M., Sun, X., Liu, Y., Chen, T., Shen, H., Zhang, Z., Zhang, H., Pan, G., Wang, W., and Xue, L.: Two-year online measurements of volatile organic compounds (VOCs) at four sites in a Chinese city: Significant impact of petrochemical industry, Sci Total Environ, 858, 159951, https://doi.org/10.1016/j.scitotenv.2022.159951, 2023.

Mukhamatdinov, I. I., Salih, I. S., Khelkhal, M. A., and Vakhin, A. V.: Application of aromatic and industrial solvents for
490 enhancing heavy oil recovery from the Ashalcha field, Energy Fuels, 35, 374-385, https://doi.org/10.1021/acs.energyfuels.0c03090, 2020.

Nelson, D., Choi, Y., Sadeghi, B., Yeganeh, A. K., Ghahremanloo, M., and Park, J.: 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,
495 https://doi.org/10.1016/j.envpol.2023.122223, 2023.

Ning, Z., Gao, S., Gu, Z., Ni, C., Fang, F., Nie, Y., Jiao, Z., and Wang, C.: Prediction and explanation for ozone variability using cross-stacked ensemble learning model, Sci. Total Environ., 935, 173382, https://doi.org/10.1016/j.scitotenv.2024.173382, 2024.

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

Pichler, M. and Hartig, F.: Machine learning and deep learning—A review for ecologists, Methods Ecol. Evol., 14, 994-1016, https://doi.org/10.1111/2041-210X.14061, 2023.

Pinthong, N., Thepanondh, S., Kultan, V., and Keawboonchu, J.: Characteristics and impact of VOCs on ozone formation potential in a petrochemical industrial area, Thailand, Atmosphere, 13, 732, https://doi.org/10.3390/atmos13050732, 2022.

505 Ragothaman, A. and Anderson, W. A.: Air quality impacts of petroleum refining and petrochemical industries, Environments, 4, 66, https://doi.org/10.3390/environments4030066, 2017.

Robin, Y., Amann, J., Baur, T., Goodarzi, P., Schultealbert, C., Schneider, T., and Schütze, A.: High-performance VOC quantification for IAQ monitoring using advanced sensor systems and deep learning, Atmosphere, 12, 1487, https://doi.org/10.3390/atmos12111487, 2021.

510 Salcedo-Sanz, S., Pérez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Pérez, D., Kadow, C., Fister, D., Barriopedro, D., García-Herrera, R., and Giuliani, M.: Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review, Theor. Appl. Climatol., 155, 1-44, https://doi.org/10.1007/s00704-023-04571-5, 2024.

Sharma, A. K., Sharma, M., Sharma, A. K., and Sharma, M.: Mapping the impact of environmental pollutants on human health and environment: A systematic review and meta-analysis, J. Geochem. Explor., 107325, https://doi.org/10.1016/j.gexplo.2023.107325, 2023.

Sharma, S., Sharma, P., and Khare, M.: Photo-chemical transport modelling of tropospheric ozone: A review, Atmos. Environ., 159, 34-54, https://doi.org/10.1016/j.atmosenv.2017.03.047, 2017.

Sharma, S., Singhal, A., Venkatramanan, V., Verma, P., and Pandey, M.: Variability in air quality, ozone formation potential by VOCs, and associated air pollution attributable health risks for Delhi's inhabitants, Environ. Sci.: Atmos., 4, 897-910, https://doi.org/10.1039/d4ea00064a, 2024.

Sillman, S.: The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments, Atmos. Environ., 33, 1821-1845, https://doi.org/10.1016/S1352-2310(98)00345-8, 1999.

Song, M., Li, X., Yang, S., Yu, X., Zhou, S., Yang, Y., Chen, S., Dong, H., Liao, K., and Chen, Q.: Spatiotemporal 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, 2021.

Tan, Y., Han, S., Chen, Y., Zhang, Z., Li, H., Li, W., Yuan, Q., Li, X., Wang, T., and Lee, S.-c.: 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, 2021.

Wang, H., Lyu, X., Guo, H., Wang, Y., Zou, S., Ling, Z., Wang, X., Jiang, F., Zeren, Y., Pan, W., Huang, X., and Shen, J.: Ozone pollution around a coastal region of South China Sea: interaction between marine and continental air, Atmos. Chem. Phys., 18, 4277-4295, https://10.5194/acp-18-4277-2018, 2018.

Wang, S., Zhao, Y., Han, Y., Li, R., Fu, H., Gao, S., Duan, Y., Zhang, L., and Chen, J.: 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, 2022.

Wang, Y., Jiang, S., Huang, L., Lu, G., Kasemsan, M., Yaluk, E. A., Liu, H., Liao, J., Bian, J., and Zhang, K.: Differences between VOCs and NOx transport contributions, their impacts on O3, and implications for O3 pollution mitigation based on CMAQ simulation over the Yangtze River Delta, China, Sci. Total Environ., 872, 162118, https://doi.org/10.1016/j.scitotenv.2023.162118, 2023.

Washenfelder, R., Trainer, M., Frost, G., Ryerson, T., Atlas, E., De Gouw, J., Flocke, F., Fried, A., Holloway, J., and Parrish, D.: Characterization of NOx, SO2, ethene, and propene from industrial emission sources in Houston, Texas, J. Geophys. Res.: Atmos., 115, https://doi.org/10.1029/2009JD013645, 2010.

Wu, Y., Fan, X., Liu, Y., Zhang, J., Wang, H., Sun, L., Fang, T., Mao, H., Hu, J., and Wu, L.: Source apportionment of VOCs based on photochemical loss in summer at a suburban site in Beijing, Atmos. Environ., 293, 119459, https://doi.org/10.1016/j.atmosenv.2022.119459, 2023.

Xiao, Z., Yang, X., Gu, H., Hu, J., Zhang, T., Chen, J., Pan, X., Xiu, G., Zhang, W., and Lin, M.: Characterization and sources of volatile organic compounds (VOCs) during 2022 summer ozone pollution control in Shanghai, China, Atmos. Environ., 327, 120464, https://doi.org/10.1016/j.atmosenv.2024.120464, 2024.

Xu, Z., Zou, Q., Jin, L., Shen, Y., Shen, J., Xu, B., Qu, F., Zhang, F., Xu, J., and Pei, X.: Characteristics and sources of ambient
550   Volatile Organic Compounds (VOCs) at a regional background site, YRD region, China: Significant influence of solvent evaporation during hot months, Sci. Total Environ., 857, 159674, https://doi.org/10.1016/j.scitotenv.2022.159674, 2023.

Yang, M., Li, F., Huang, C., Tong, L., Dai, X., and Xiao, H.: VOC characteristics and their source apportionment in a coastal industrial area in the Yangtze River Delta, China, J. Environ. Sci., 127, 483-494, https://doi.org/10.1016/j.jes.2022.05.041, 2023.

555   Yao, D., Tang, G., Wang, Y., Yang, Y., Wang, L., Chen, T., He, H., and Wang, Y.: Significant contribution of spring northwest transport to volatile organic compounds in Beijing, J. Environ. Sci., 104, 169-181, https://doi.org/10.1016/j.jes.2020.11.023, 2021.

Zhang, M., Liu, Y., Xu, X., He, J., Ji, D., Qu, K., Xu, Y., Cong, C., and Wang, Y.: A Systematic Review on Atmospheric Ozone Pollution in a Typical Peninsula Region of North China: Formation Mechanism, Spatiotemporal Distribution, Source
560   Apportionment, and Health and Ecological Effects, Curr Pollution Rep, 11, 9, https://10.1007/s40726-024-00338-2, 2025.

Zhang, Y., Xue, L., Carter, W. P., Pei, C., Chen, T., Mu, J., Wang, Y., Zhang, Q., and Wang, W.: Development of ozone reactivity scales for volatile organic compounds in a Chinese megacity, Atmos. Chem. Phys., 21, 11053-11068, https://doi.org/10.5194/acp-21-11053-2021, 2021.

Zhang, Z., Xu, J., Ye, T., Chen, L., Chen, H., and Yao, J.: Distributions and temporal changes of benzene, toluene, ethylbenzene,
565   and xylene concentrations in newly decorated rooms in southeastern China, and the health risks posed, Atmos. Environ., 246, https://doi.org/10.1016/j.atmosenv.2020.118071, 2021.

Zhao, D., Xin, J., Wang, W., Jia, D., Wang, Z., Xiao, H., Liu, C., Zhou, J., Tong, L., and Ma, Y.: Effects of the sea-land breeze on coastal ozone pollution in the Yangtze River Delta, China, Sci. Total Environ., 807, 150306, https://doi.org/10.1016/j.scitotenv.2021.150306, 2022