



Multi-machine-learning approaches to modeling small-scale source attribution of ozone formation

Zheng Xiao^{1,2}, Yifeng Lu³, and Guangli Xiu^{1,2}

¹State 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, China

²Shanghai 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, China

³Shanghai Chemical Industry Park Administration Committee, Shanghai 201507, China

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

Received: 14 January 2025 – Discussion started: 5 March 2025

Revised: 28 June 2025 – Accepted: 9 July 2025 – Published: 8 May 2026

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 ($R^2 = 0.644$) 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, VOC source identification technology mainly includes source emission inventories (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). However, it should be emphasized that PMF and CTMs are methodologically distinct: CTMs simulate physico-chemical processes driven by emission inventories and meteorology, enabling dynamic regional-scale predictions, whereas PMF statistically decomposes sources based on covariance in observed species concentrations. While PMF demonstrates strong performance in settings with spatially dense monitoring, its accuracy is inherently constrained by input data representativeness and chemical stability of source profiles. Consequently, PMF is best positioned as a complementary tool for CTMs in integrated assessment frameworks. PMF models are typically combined with the O₃ formation potential (OFP) calculated by the maximum incremental reactivity (MIR) value of VOC species (Carter, 2010) to assess the relationship between VOCs and O₃ 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 O₃ exceeds the standard, whether the MIR value can fully reflect the contribution of VOCs to O₃ under complex atmospheric pollution conditions in China is controversial (Zhang et al., 2021). Furthermore, conventional methods like PMF, while effective for source categorization, face critical limitations: they cannot differentiate rapid O₃ formation via photostationary state perturbations (e.g., alkene depletion) from slower HO_x-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 O₃ 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; Lundberg and Lee, 2017). Novel ML approaches can provide robust predictions with less reliance on mechanistic details, making them attractive alternatives or comple-

mentary 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 O₃ 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. (2023) assessed the impact of emission sources on O₃ formation by combining PMF with four ML models. Chen et al. (2024) used a PMF model combined with the Category Boosting (CatBoost) model and the Shapley additive interpretation algorithm to quantify the influence of pollution sources and meteorological factors on VOCs. Ning et al. (2024) constructed a cross-stacked ensemble learning model (CSEM) to predict O₃ concentrations under different NO_x and VOC 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 VOCs (e.g., propylene, 1,3-butadiene) drive rapid O₃ formation via alkene-NO_x reactions within < 50 km under sea-land breezes (3–5 m s⁻¹) (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, highly reactive VOCs (e.g., propylene, 1,3-butadiene; atmospheric lifetime < 6 h) that dominate O₃ 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 O₃ formation with unprecedented spatial resolution. Our methodology 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 assess VOC–O₃ 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 O₃ 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 O₃ formation. The approach transcends traditional analytical limitations by offering a powerful tool for high-resolution source traceability, thereby providing crucial support for implementing targeted and effective O₃ 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 in the Supplement). The distribution of these monitoring stations is illustrated in Fig. 1. Hourly O₃ concentration data utilized in this study were obtained from the Atmospheric Environmental Monitoring Routine Data Management System of the Shanghai Environmental Monitoring Station (<https://github.com/boboob/Multi-machine-learning-approaches>, last access: 29 April 2026). 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://github.com/boboob/Multi-machine-learning-approaches>, last access: 29 April 2026). The dataset covers a collection period from 1 January 2021 to 31 December 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 enabling 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^{−1} for the initial 10 min 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 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.

2.2 Positive matrix factorization model (PMF)

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 (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 objective of minimizing the discrepancy between the calculated *Q* value and the theoretical *Q* value. The formulation of the PMF model is repre-

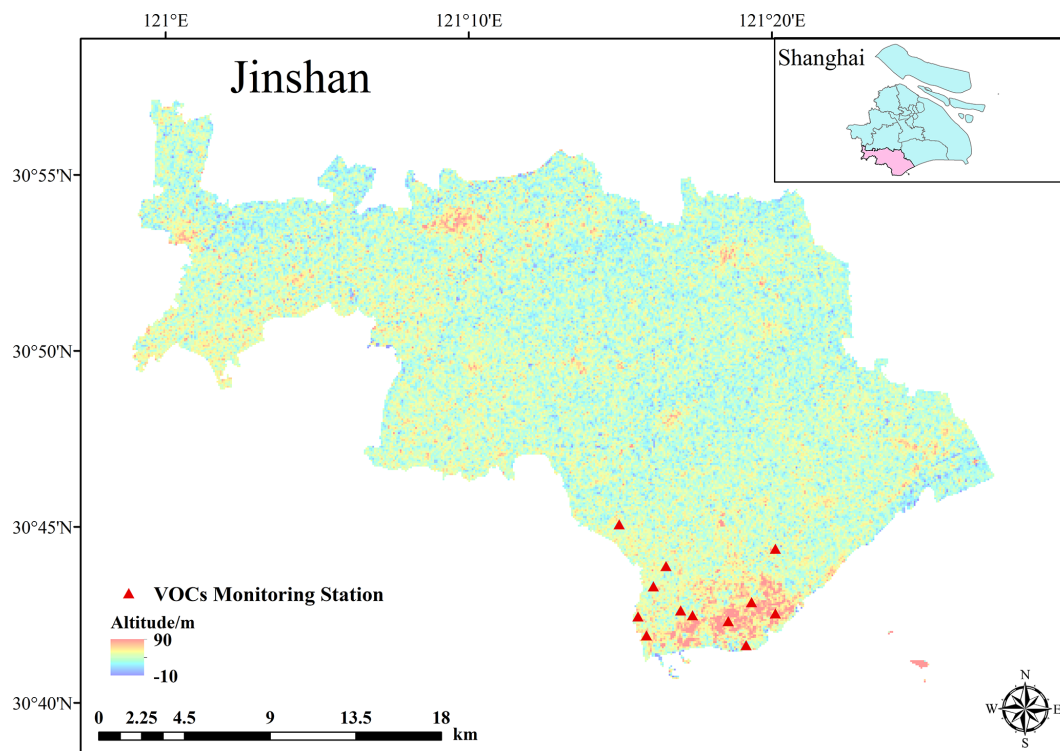


Figure 1. Study area and sampling sites. The triangles indicate the 12 sites utilized for the evaluation of the model.

sented by Eq. (1):

$$e_{ij} = x_{ij} - \sum_{k=1}^p g_{ik} f_{kj}, \quad (1)$$

where X_{ij} represents the concentration of j th species in the i th sample, g_{ik} represents the concentration of k th source in the i th sample, f_{kj} represents the mass percentage of j th species in the k th sample, and e_{ij} represents the residue factor for j th species in the i th sample.

In the PMF model, factor contributions and source fingerprint spectra are derived by minimizing the objective function (Q), as described in Eq. (2):

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

The uncertainty (u_{ij}) of the j th species in the i th sample is calculated by considering the minimum detection limit (MDL) and error scores of each species.

Crucially, PMF resolves source-specific contributions to VOC mass concentrations (\mathbf{g} matrix), not directly to ozone formation. To attribute ozone impacts, we integrate the PMF-derived \mathbf{g} matrix as inputs to machine learning models.

2.3 Machine learning models

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

ering the period from January 2021 to December 2023 for training and evaluating ML models. In this study, hourly data of O₃ and total volatile organic compound (TVOC) concentrations from 12 monitoring stations were input into the ML models. The dataset was partitioned into training (70 %) and testing (30 %) subsets using chronological splitting to preserve temporal integrity. Model robustness was further validated via 10-fold cross-validation on the training set. Subsequently, the ML models were integrated with the SHAP (SHapley Additive exPlanations) (Sect. S1.2 for model details) framework to obtain SHAP values for the 12 stations, quantifying their contributions to O₃ formation. SHAP values quantify the relative influence of input features (e.g., site-specific TVOC concentrations) on ML-predicted O₃ levels – not absolute physicochemical contributions to ozone formation. Higher absolute SHAP magnitudes indicate stronger feature importance in the model's decision-making, enabling spatial prioritization of emission hotspots. The stations contributing the most to O₃ 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 O₃ concentration data, were then input into the ML models for SHAP analysis. Finally, the optimal ML model and the most severe pollution sources were identified. The source contribution matrix (\mathbf{g} matrix) derived from PMF analysis, representing the hourly contribution rates (%) of the six resolved sources, served as input features alongside concurrent O₃ concentration data for training the machine

learning models. This approach transformed source apportionment results into predictive variables for O₃ formation modeling. Subsequently, the SHAP algorithm was applied to quantify the contribution of each PMF-resolved source to O₃ predictions generated by the optimized ML model, enabling high-resolution attribution of emission sources to ozone formation dynamics.

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 Sect. S1.1 for model details). Compared to complex deep learning architectures (e.g., CNNs, RNNs), these models offer distinct advantages for high-dimensional spatiotemporal datasets: (1) native support for feature importance metrics (e.g., Gini, permutation importance) enables direct interpretation of predictor contributions without post hoc explainers, (2) computational efficiency facilitates rigorous hyperparameter optimization with Bayesian methods and (3) robustness to collinear features common in atmospheric chemistry (Pichler and Hartig, 2023; Kaur et al., 2020). Bayesian optimization (Robin et al., 2021) was then applied to determine optimal hyperparameters across more than 120 000 hourly samples. Detailed implementation procedures of Bayesian optimization, including model-specific hyperparameter spaces, convergence criteria and computational configurations, are comprehensively documented in Sect. S1.3 of the Supplement. Ultimately, through iterative optimization of training configurations (70 : 30 train–test partitioning, 10-fold cross-validation) and six repetitions of randomized initialization experiments with distinct random seeds (Seed = 42, 87, 124, 256, 512, 1024), we selected the model with the highest R^2 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.

R^2 is an indicator that measures the overall goodness of fit of a regression model, and its calculation is shown in Eq. (3):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3)$$

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. (4):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4)$$

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. (5):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (5)$$

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

3 Results and discussions

3.1 Spatiotemporal characterization of ozone and VOC 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 O₃ concentrations across multiple monitoring sites from 2021 to 2023 reveals distinct patterns that inform our ML-based source attribution approach. The temporal evolution of O₃ concentrations shows a notable progression, with mean values of 64.45, 68.79 and 68.90 $\mu\text{g m}^{-3}$ recorded in 2021, 2022 and 2023, respectively (Fig. 3a). The relatively lower O₃ 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

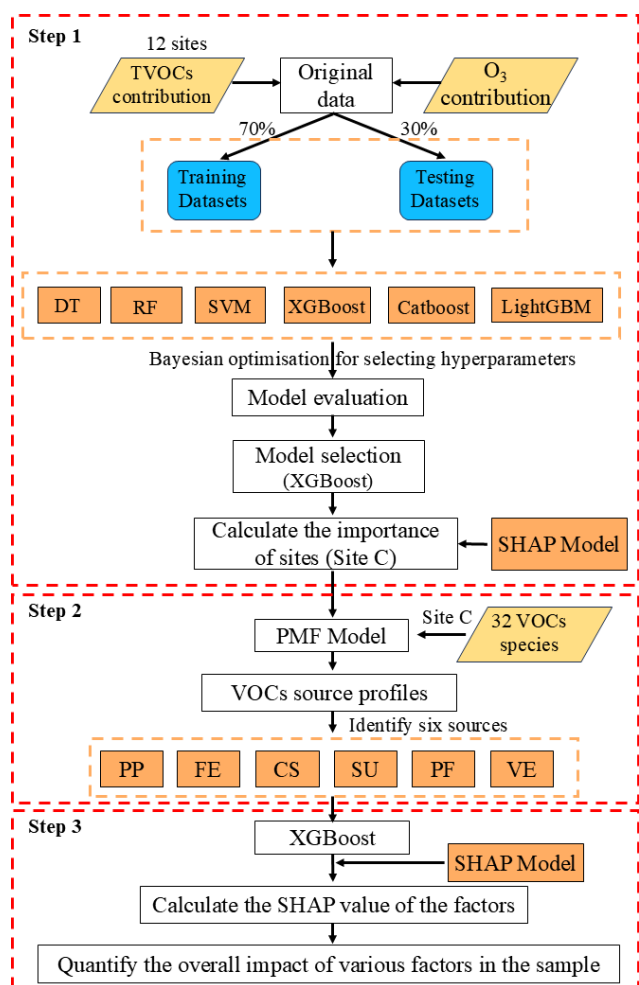


Figure 2. Schematic workflow of the integrated ML–PMF framework for ozone source attribution (see Sect. 2.3 for methodological details).

2022, primarily attributed to the implementation of stringent emission control measures by local regulatory authorities (Xiao et al., 2024). The seasonal analysis reveals distinctive O₃ formation patterns in the study area (Fig. 3b). O₃ 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 O₃ 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. O₃ concentrations follow a characteristic curve, initiating a rapid increase at 08:00 (UTC+8), sustain-

ing 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 O₃ 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 O₃ 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.

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 10-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 R^2 value of 0.637.

Further validation through scatter plot analysis (Fig. S2) revealed that the XGBoost model exhibited significantly smaller deviations between predicted and observed values compared to other algorithms, confirming its enhanced predictive accuracy. Based on these comprehensive evaluation

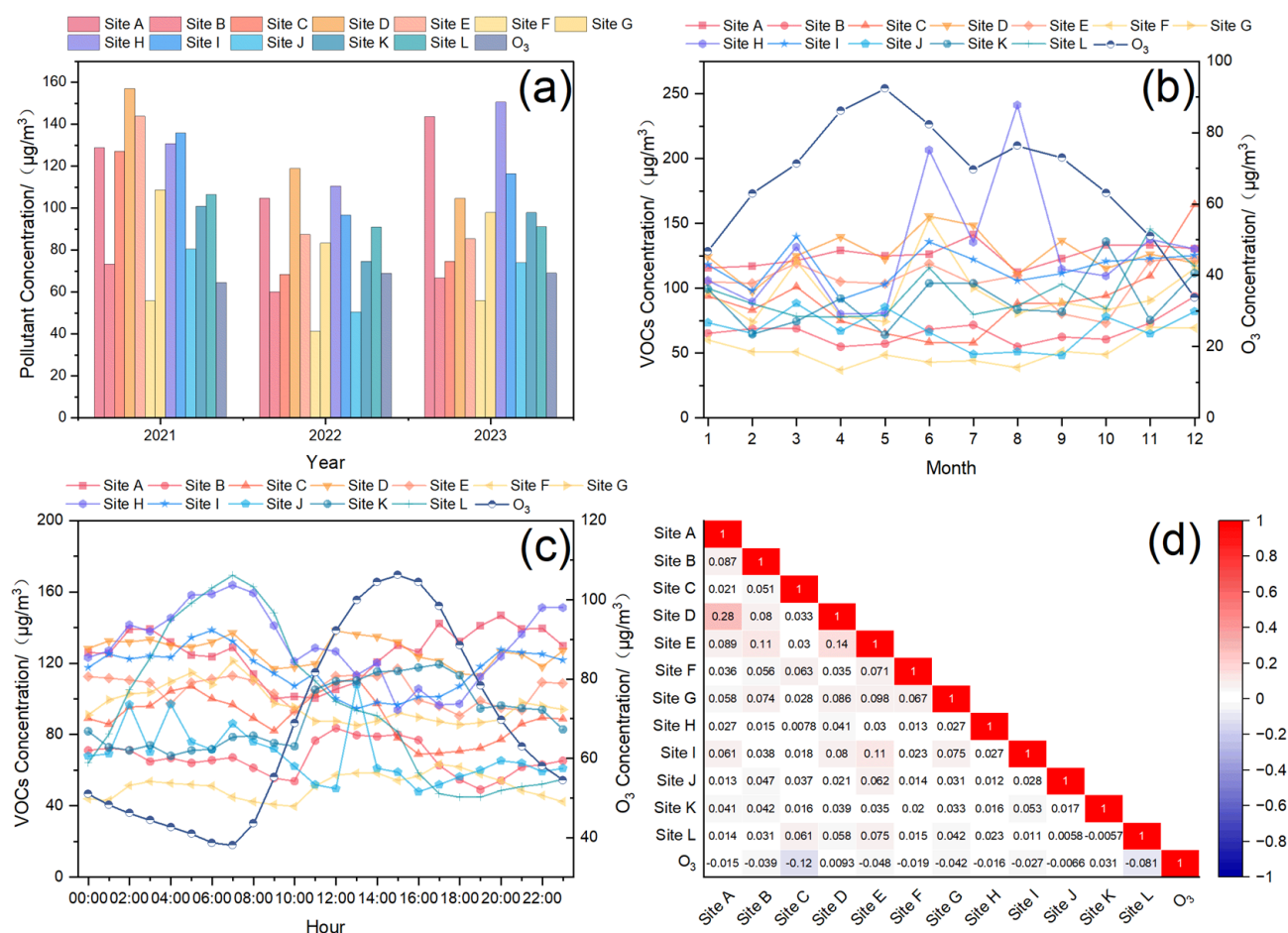


Figure 3. Temporal and spatial dynamics of VOCs and O₃, along with their correlations, as observed at 12 monitoring sites. **(a)** Mean concentration variations of O₃ and VOCs from 2021 to 2023. **(b)** Monthly profiles of O₃ and VOC concentrations. **(c)** Daily profiles of O₃ and VOC concentrations.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

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 of source attribution results.

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 XGBoost regression model (Fig. 4). The analysis revealed 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. This SHAP-negative

TVOC signal reflects the model's recognition of Site C's unique emission mix, where VOC-rich plumes initially suppress ozone via NO titration under proximate NO_x-rich conditions but subsequently fuel downwind ozone formation as air masses age and transition to VOC-limited chemical regimes. This duality aligns with observational studies in petrochemical clusters, where proximal VOC hotspots exhibit transient O₃ suppression due to localized precursor interactions while driving net regional ozone increases through secondary photochemical pathways (Guo et al., 2022; Ren et al., 2024). 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 Sect. 3.1.

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 fa-

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

Model	10-fold cross-validation (R^2)	R^2	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, RF: Random Forest, SVM: Support Vector Machine, XGBoost: Extreme Gradient Boosting, CatBoost: Category Boosting, LightGBM: Light Gradient Boosting Machine

cilities 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 emissions.

To validate the robustness of these findings, we extended the SHAP analysis to encompass five additional ML models (Fig. S3). 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 patterns. Importantly, SHAP values here reflect the relative importance of each site's TVOC concentrations to XGBoost-predicted O₃ – not direct source contributions. This approach identifies locations where VOC variations most strongly perturb O₃ predictions (e.g., Site C's dominant role), guiding subsequent PMF-based source apportionment.

In addition, to ensure robustness of feature importance interpretations, we employed three complementary attribution schemes beyond SHAP: (1) SAGE (Shapley Additive Global importance) for global feature relevance; (2) Gini importance for intrinsic tree-based rankings and (3) permutation importance evaluating prediction degradation under feature shuffling. Figure S6 compares normalized importance scores across SHAP, SAGE, Gini and permutation methods for the sites. For Site C, SHAP, SAGE and permutation importance uniformly assign it the highest score (normalized score ≈ 1.0), confirming its dominance in prediction-sensitivity-based frameworks. In contrast, Gini importance ranks Site C third (normalized score ≈ 0.6). To further quantify the consistency of feature importance rankings across methodologies, we computed pairwise Pearson correlation coefficients between SHAP, SAGE, Gini and permutation importance scores (Fig. S7). The correlation matrix reveals near-perfect agreement between SHAP, SAGE and permutation importance (Pearson $r = 0.98$ for all inter-method pairs), confirming these techniques capture overlapping dimensions of feature relevance tied to global prediction sen-

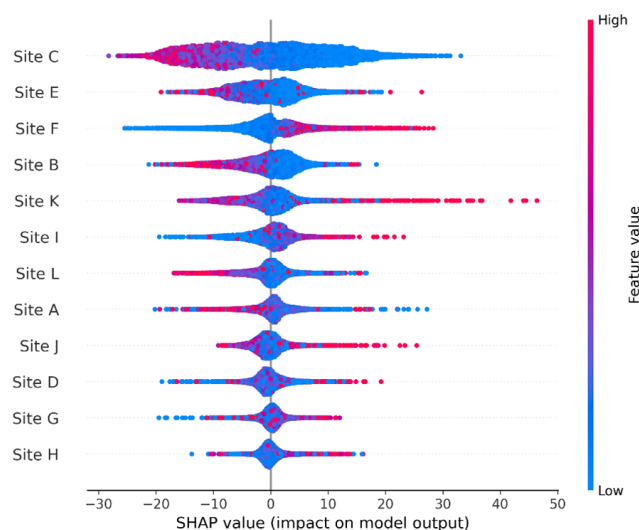


Figure 4. Feature importance of total volatile organic compound (TVOC) drivers obtained by XGBoost model. Blue points indicate negative contributions to the prediction, while red points represent positive contributions.

sitivity. In contrast, Gini importance exhibits weak correlation with the other three schemes ($r \leq 0.12$). This divergence arises because Gini importance prioritizes local decision-tree split purity (e.g., maximizing variance reduction at individual nodes), which can overemphasize features with high within-tree variability but low global relevance to ozone photochemistry. Conversely, SHAP, SAGE and permutation – grounded in global prediction sensitivity – robustly identify Site C as the primary driver of O₃ formation, as their concordance ($r = 0.98$) reflects shared sensitivity to chemically meaningful VOC–O₃ interactions. Thus, the methodological triangulation of SHAP, SAGE and permutation – coupled with Gini's demonstrated insensitivity to global photochemical dynamics – unequivocally confirms Site C's preeminent role in regional ozone formation.

3.3 Source apportionment and temporal–spatial characterization of VOCs based on the PMF model

3.3.1 Analytical framework and results of the PMF model

Inter-species correlation analysis revealed distinct clustering patterns among VOCs (Fig. S4). Strong positive correlations ($r > 0.7$) were observed among alkane homologues such as ethane (C₂), propane (C₃), *n*-butane (C₄) and isopentane (C₅) ($r = 0.71$ – 0.92), indicative of co-emission from fuel evaporation and petrochemical activities (Yang et al., 2024). Strong positive correlations were observed between propylene and 1,3-butadiene ($r = 0.82$), indicative of shared emission pathways from petrochemical cracking processes (Zhou et al., 2021; White, 2007). In addition, aromatic compounds including toluene, *m*-/*p*-xylene and ethylbenzene formed a highly correlated group ($r = 0.75$ – 0.95), frequently used in solvent applications and industrial coatings (Weiss, 1997). Chlorinated VOCs such as 1,2-dichloroethane and monochlorobenzene exhibit moderate correlations ($r = 0.44$), suggesting mixed contributions from polymer manufacturing and solvent use (Huang et al., 2014). Conversely, propane displayed weak correlations with aromatic species ($r < 0.2$), aligning with its dominance in combustion sources. The observed clustering can provide a basis for PMF-derived source profiles, as the species of interest are predominantly aligned with common emission processes.

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, and (2) highly reactive compounds were omitted unless serving as specific tracers for particular sources. Ultimately, 36 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.

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 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 (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 4 reached 69.12 %. Currently, aromatic hydrocarbons are widely used as solvents in industrial production (Zhang et al., 2021; 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, 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 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.

3.3.2 Seasonal variation and contributing factors of VOC source distribution

Seasonal analysis unveiled significant temporal variations in source contributions (Fig. 6). 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

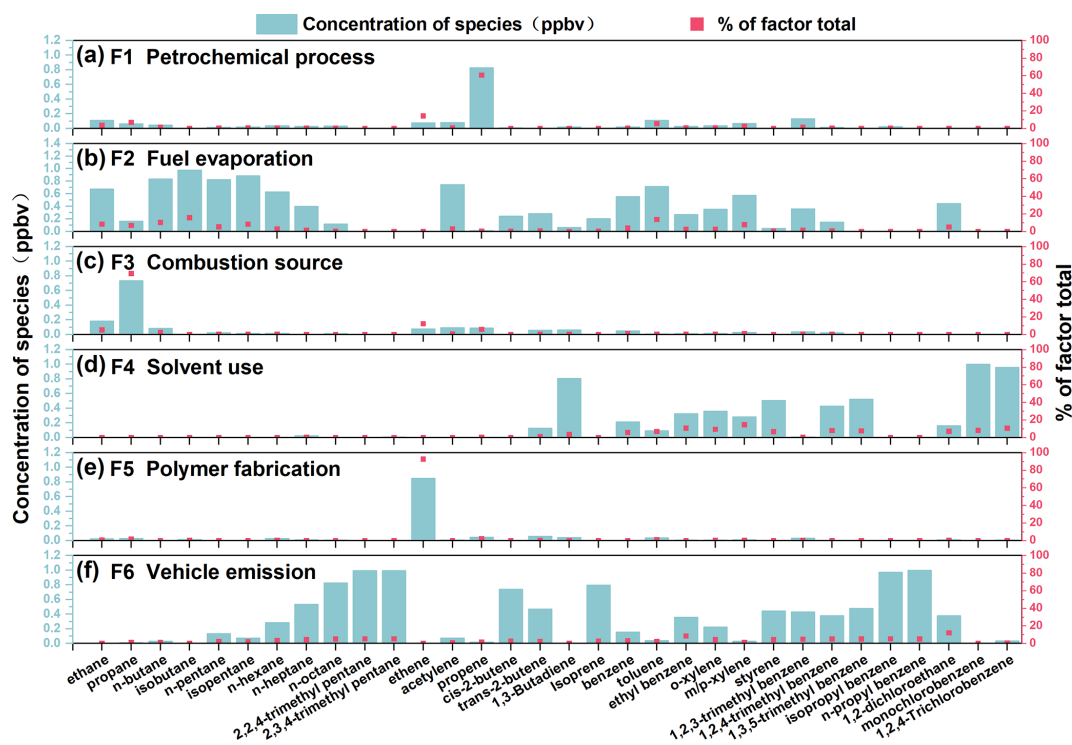


Figure 5. Source profiles calculated using the PMF model in site C.

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 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, subsequently promoting O₃ 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).

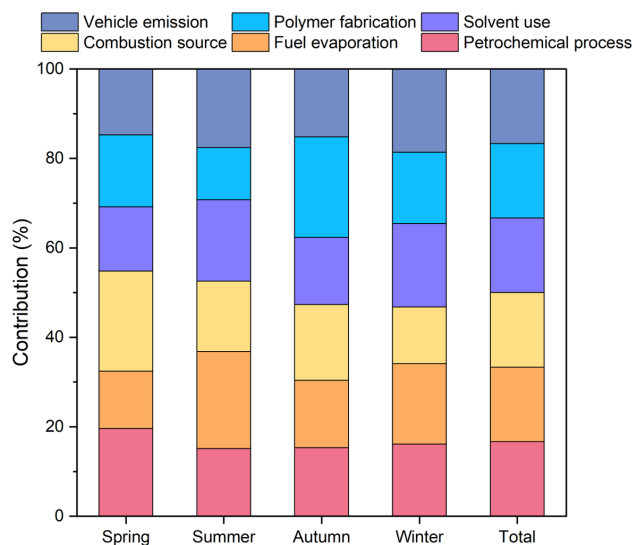


Figure 6. Influence of various sources on atmospheric VOCs across different seasons.

3.4 Quantitative assessment of source-specific contributions to O₃ formation

The integrated PMF-SHAP framework was employed to quantitatively evaluate the source-specific contributions to O₃ formation in the Jinshan industrial complex during 2021–2023. Note that SHAP-derived source contributions are de-

rived from the PMF-resolved source profiles (**g** matrix), not site-level TVOC data. This integrated approach directly links emission sources to O₃ impacts. Through the coupling of XGBoost-SHAP modeling with corresponding SHAP values, we systematically assessed the relative importance of various emission sources in driving O₃ pollution dynamics (Fig. 7). The analysis revealed that SU and FE were the predominant contributors to O₃ formation, 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 O₃ 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 O₃ pollution control strategies within industrial contexts.

To quantify the contribution of VOC sources identified by PMF to O₃ formation, we integrated the mass contributions of VOCs from specific sources with their respective maximum incremental reactivity (MIR) values (Carter, 2010; Venecek et al., 2018). Subsequently, we employed ozone formation potential (OFP) to assess the relative contributions of different VOCs to O₃ generation. The OFP quantification of PMF-resolved sources (Fig. S5) aligns robustly with SHAP-derived source prioritization, validating the scientific coherence of the PMF framework. SU exhibits the highest OFP contribution (32.20 μg m⁻³), driven by its dominant aromatic constituents (*m*-/*p*-xylene: 14.55 %, MIR = 5.47 gO₃/gVOC; ethyl benzene: 10.57 %, MIR = 3.11 gO₃/gVOC; *o*-xylene: 9.28 %, MIR = 7.17 gO₃/gVOC) whose combined reactivity (MIR-weighted OFP = 14.85 μg m⁻³) amplifies its ozone-driving potential despite moderate mass abundance (34.40 % of total VOCs). In contrast, FE (26.3 μg m⁻³) demonstrated a higher OFP, driven by the substantial mass proportion of C₂–C₅ hydrocarbons (68.15 %) and their high mean concentration (11.81 μg m⁻³). The SHAP value (3.00) confirmed its disproportionate influence on O₃ formation relative to its mass contribution. Additionally, VE (10.10 μg m⁻³) exhibited the lowest OFP, primarily due to the low reactivity of 1,2-dichloroethane (12.50 % by mass, MIR = 0.23). The SHAP value (1.92) reflected its limited photochemical impact. To contextualize the ML-derived findings within established regional photochemistry, we explicitly reconcile our SHAP-based source rankings (SU > FE > CS) with empirical kinetic modeling approach (EKMA) studies specific to the Shanghai. The dominance of solvent use (SU) and fuel evaporation (FE) sources aligns with EKMA analyses demonstrating VOC-limited regimes in Shanghai's industrial corridors (Zhang et al., 2024), where reactive aromatics (*m*-/*p*-xylene) and short-chain alkenes (propylene, 1,3-butadiene) drive > 70 % of incremental reactivity.

The seasonal decomposition of source contributions through SHAP analysis demonstrated strong concordance with PMF-derived temporal patterns (Fig. 8). CS emerged

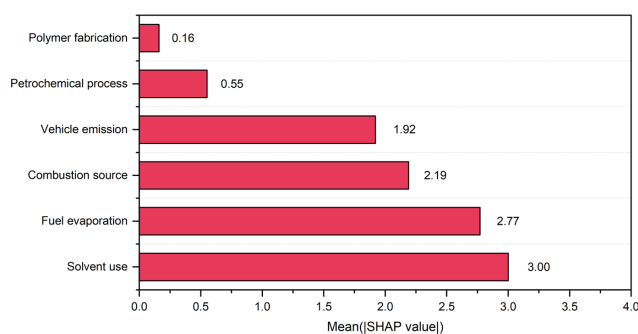


Figure 7. Mean SHAP values for six pollution sources affecting ozone levels.

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 O₃ management strategies.

Contrary to conventional CTM-based studies emphasizing combustion sources (CS) as primary O₃ drivers, our ML-PMF-SHAP integration identifies solvent use (SU, SHAP = 3.00) and fuel evaporation (FE, SHAP = 2.77) as dominant contributors. This shift highlights industrial process-specific emissions (e.g., aromatic solvents, light alkanes) as critical O₃ precursors in petrochemical zones, challenging broad regional assumptions. Site C exhibits a unique negative SHAP-O₃ correlation despite high VOC loads. We attribute this to rapid NO titration from proximate NO_x-rich plumes (e.g., gas stations, port activities), suppressing local O₃ while fueling downwind formation (> 5 km) as air masses age into VOC-limited regimes – a micro-scale dynamic unresolvable by traditional PMF-OFP methods. SHAP quantifies disproportionate O₃ impacts unconstrained by VOC mass. For example, SU contributes only 34.40 % of total VOCs but drives 32.20 μg m⁻³ OFP due to high-reactivity aromatics (e.g., *o*-xylene, MIR = 7.17 gO₃/gVOC), whereas FE's higher mass (68.15 % alkanes) yields lower OFP (26.3 μg m⁻³) but amplified SHAP influence via spatial persistence. These insights establish a paradigm for facility-scale O₃ control, prioritizing SU/FE reductions over combustion sources in coastal industrial clusters. 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-O₃ periods, would optimize the effectiveness of O₃ pollution mitigation efforts in industrial areas.

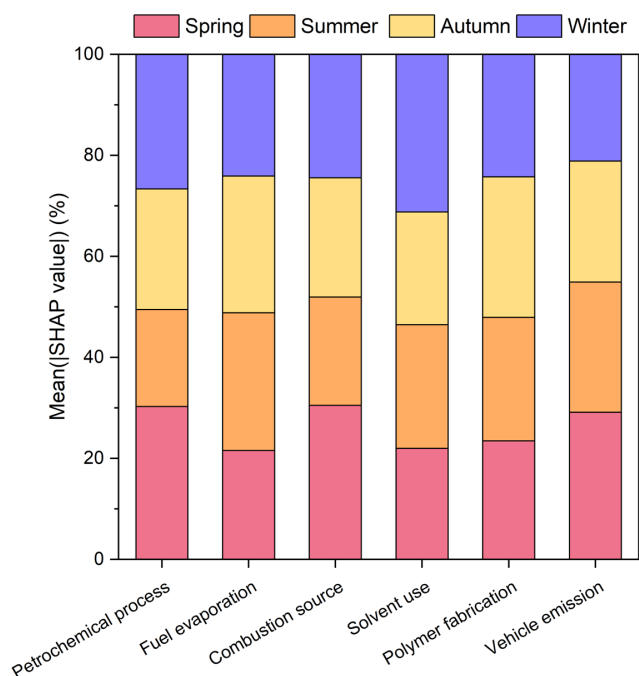


Figure 8. Proportional contributions of six pollution sources to ozone levels, represented by SHAP values across different seasons.

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 tar-

geted 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. While our integrated PMF-ML-SHAP framework advances high-resolution source attribution for industrial O₃ formation, three key limitations warrant consideration: the 1 km × 1 km spatial resolution, though unprecedented for facility-scale analysis, may not resolve sub-facility emission hotspots (e.g., individual storage tanks or pipeline leaks), necessitating future integration of drone-based hyperspectral sensors or stack-level monitors for hyperlocal validation. Furthermore, reliance on hourly GC-FID measurements could underestimate fast-reacting alkenes (e.g., propylene, isoprene) with atmospheric lifetimes < 2 h; real-time proton-transfer-reaction mass spectrometry (PTR-MS) would enhance temporal resolution for such compounds. Lastly, while optimized for Shanghai's coastal petrochemical environment – characterized by high humidity (> 75 %), sea–land breezes (3–5 m s^{−1}) and reactive aromatic–alkene mixtures – the framework's efficacy requires validation in inland/arid industrial regions with distinct meteorology (e.g., lower humidity, weaker advection) and emission profiles (e.g., solvent-dominated manufacturing clusters). These constraints highlight inherent trade-offs between resolution and practicality but do not invalidate the framework's utility for targeted O₃ management in complex industrial zones. 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.

Appendix A: Acronym glossary

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

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, are confidential. Hourly O₃ concentration data utilized in this study were obtained from the Atmospheric Environmental Monitoring Routine Data Management System of the Shanghai Environmental Monitoring Station: <https://github.com/boboob/Multi-machine-learning-approaches> (last access: 29 April 2026). 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://github.com/boboob/Multi-machine-learning-approaches> (last access: 29 April 2026).

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/acp-26-6117-2026-supplement>.

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.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

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 grants from the National Ministry of Science and Technology (grant nos. 2022YFC3703501 and 2022YFC3703503) and the Shanghai Jinshan Municipal Bureau of Ecology and Environment (grant no. Huhuanke-2025-10).

Review statement. This paper was edited by Rob MacKenzie and reviewed by two anonymous referees.

References

- Baklanov, A. and Korsholm, U.: On-line integrated meteorological and chemical transport modelling: advantages and perspectives, *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*, Wiley, 61–103, <https://doi.org/10.1002/9781119159797>, 2017.
- 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.
- 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.
- 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, <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 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.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Gian-notti, 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.
- Guo, W., Yang, Y., Chen, Q., Zhu, Y., Zhang, Y., Zhang, Y., Liu, Y., Li, G., Sun, W., and She, J.: Chemical reactivity of volatile organic compounds and their effects on ozone formation in a petrochemical industrial area of Lanzhou, Western China, *Sci. Total Environ.*, 839, 155901, <https://doi.org/10.1016/j.scitotenv.2022.155901>, 2022.
- 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-PM_{2.5} compound pollution in the Yangtze River Delta, *Sci. Total Environ.*, 906, 167503, <https://doi.org/10.1016/j.scitotenv.2023.167503>, 2024.
- Huang, B., Lei, C., Wei, C., and Zeng, G.: Chlorinated volatile organic compounds (Cl-VOCs) in environment – sources, potential human health impacts, and current remediation technologies, *Environ. Int.*, 71, 118–138, <https://doi.org/10.1016/j.envint.2014.06.013>, 2014.
- 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, 117340, <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*, 2020, Honolulu, HI, USA, 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 NO_x and VOC emissions: Examining machine learning models, *Environ. Int.*, 176, 107969, <https://doi.org/10.1016/j.envint.2023.107969>, 2023.
- Li, M., Zhang, Q., Streets, D. G., He, K. B., Cheng, Y. F., Emmons, L. K., Huo, H., Kang, S. C., Lu, Z., Shao, M., Su, H., Yu, X., and Zhang, Y.: 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.
- 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.
- 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.: PM_{2.5} 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-NO_x-VOCs relationships induced by anomalous ozone in industrial regions during COVID-19 in Shanghai, *Atmos. Environ.*, 308, 119864, <https://doi.org/10.1016/j.atmosenv.2023.119864>, 2023.
- 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.
- Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, *arXiv [preprint]*, 30, <https://doi.org/10.48550/arXiv.1705.07874>, 2017.
- 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, <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 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, <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, [https://doi.org/10.1016/S0169-7439\(96\)00044-5](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.
- 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.
- Ren, H., Xia, Z., Yao, L., Qin, G., Zhang, Y., Xu, H., Wang, Z., and Cheng, J.: Investigation on ozone formation mechanism and control strategy of VOCs in petrochemical region: insights from chemical reactivity and photochemical loss, *Sci. Total Environ.*, 914, 169891, <https://doi.org/10.1016/j.scitotenv.2024.169891>, 2024.
- 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.
- 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.*, 255, 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, NO_x and hydrocarbons in urban and polluted rural environments, *Atmos. Environ.*, 33, 1821–1845, [https://doi.org/10.1016/S1352-2310\(98\)00345-8](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., Chen, Q., Lu, K., Zhang, N., Cao, J., Zeng, L., and Zhang, Y.: 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.
- Venecek, M. A., Carter, W. P., and Kleeman, M. J.: Updating the SAPRC Maximum Incremental Reactivity (MIR) scale for the United States from 1988 to 2010, *J. Air Waste Manage. Assoc.*, 68, 1301–1316, <https://doi.org/10.1080/10962247.2018.1498410>, 2018.
- 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://doi.org/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 NO_x 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 NO_x, SO₂, ethene, and propene from industrial emission sources in Houston, Texas, *J. Geophys. Res.-Atmos.*, 115, D16311, <https://doi.org/10.1029/2009JD013645>, 2010.
- Weiss, K. D.: Paint and coatings: A mature industry in transition, *Prog. Polym. Sci.*, 22, 203–245, [https://doi.org/10.1016/S0079-6700\(96\)00019-6](https://doi.org/10.1016/S0079-6700(96)00019-6), 1997.
- White, W. C.: Butadiene production process overview, *Chem. Biol. Interact.*, 166, 10–14, <https://doi.org/10.1016/j.cbi.2007.01.009>, 2007.
- 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 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.
- Yang, Y., Meng, X., Chen, Q., Xue, Q., Wang, L., Sun, J., Guo, W., Tao, H., Yang, L., and Chen, F.: Characteristics of volatile organic compounds under different operating conditions in a petrochemical industrial zone and their effects on ozone formation, *Environ. Pollut.*, 363, 125254, <https://doi.org/10.1016/j.envpol.2024.125254>, 2024.
- 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 Apportionment, and Health and Ecological Effects, *Curr. Pollution Rep.*, 11, 9, <https://doi.org/10.1007/s40726-024-00338-2>, 2025.
- Zhang, Y., Xue, L., Carter, W. P. L., 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, Y., Fu, Q., Wang, T., Huo, J., Cui, H., Mu, J., Tan, Y., Chen, T., Shen, H., and Li, Q.: A quantitative analysis of causes for increasing ozone pollution in Shanghai during the 2022 lockdown and implications for control policy, *Atmos. Environ.*, 326, 120469, <https://doi.org/10.1016/j.atmosenv.2024.120469>, 2024.
- Zhang, Z., Xu, J., Ye, T., Chen, L., Chen, H., and Yao, J.: Distributions and temporal changes of benzene, toluene, ethylbenzene, and xylene concentrations in newly decorated rooms in southeastern China, and the health risks posed, *Atmos. Environ.*, 246, 118071, <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.
- Zhou, X., Sun, Z., Yan, H., Feng, X., Zhao, H., Liu, Y., Chen, X., and Yang, C.: Produce petrochemicals directly from crude oil catalytic cracking, a techno-economic analysis and life cycle society-environment assessment, *J. Cleaner Prod.*, 308, 127283, <https://doi.org/10.1016/j.jclepro.2021.127283>, 2021.