Authors' responses (EGUSPHERE-2025-160)

Dear Editor and Referees:

On behalf of all the authors, I would like to express our great appreciations for your kind help and constructive comments concerning our article entitled " Multi-Machine Learning Approaches to Modeling Small-Scale Source Attribution of Ozone Formation " (Manuscript No. EGUSPHERE-2025-160). Those comments are all valuable and helpful for improving our article. According to your comments, we have made carefully revision of our manuscript to make our results convincing. In this revised vision, revisions were all highlighted within the document by using red colored text. Point-by-point responses to the editor and reviewers are listed below this letter.

Guangli Xiu

June 28, 2025

Referee #1

Xiao et al. utilised observational datasets, PMF analysis, and set of ML models to understand the importance of different emission sources of VOCs in ozone formation of ozone in Jinshan region of Shanghai. The application of novel ML approaches, particularly combining that with comprehensive VOC measurements and PMF, can be valuable. However, the paper is highly technical, with limited scientific discussions, new inferences, and evaluation of ML based results in terms of scientific understanding. Additionally, technical details connecting PMF output with ML are also not very clear. The flow of discussion, from one section to the next, needs improvement throughout the manuscript. Therefore, the manuscript, in its present form, may not be recommended for publication in ACP. Some additional details are also provided below.

1.The authors have considered > 30 different VOCs. How are these different VOCs inter-correlated?

Response: We appreciate the reviewer's constructive feedback. To address the inter-correlation analysis of VOCs, we conducted Pearson correlation analysis and incorporated the results into Section 3.3.1 (Line 286). This addition clarifies the relationships between key VOC species and their implications for source identification.

Lines 318-328: "Inter-species correlation analysis revealed distinct clustering patterns among VOCs (Fig. S4). Strong positive correlations ($r > 0.7$) 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), which are 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."

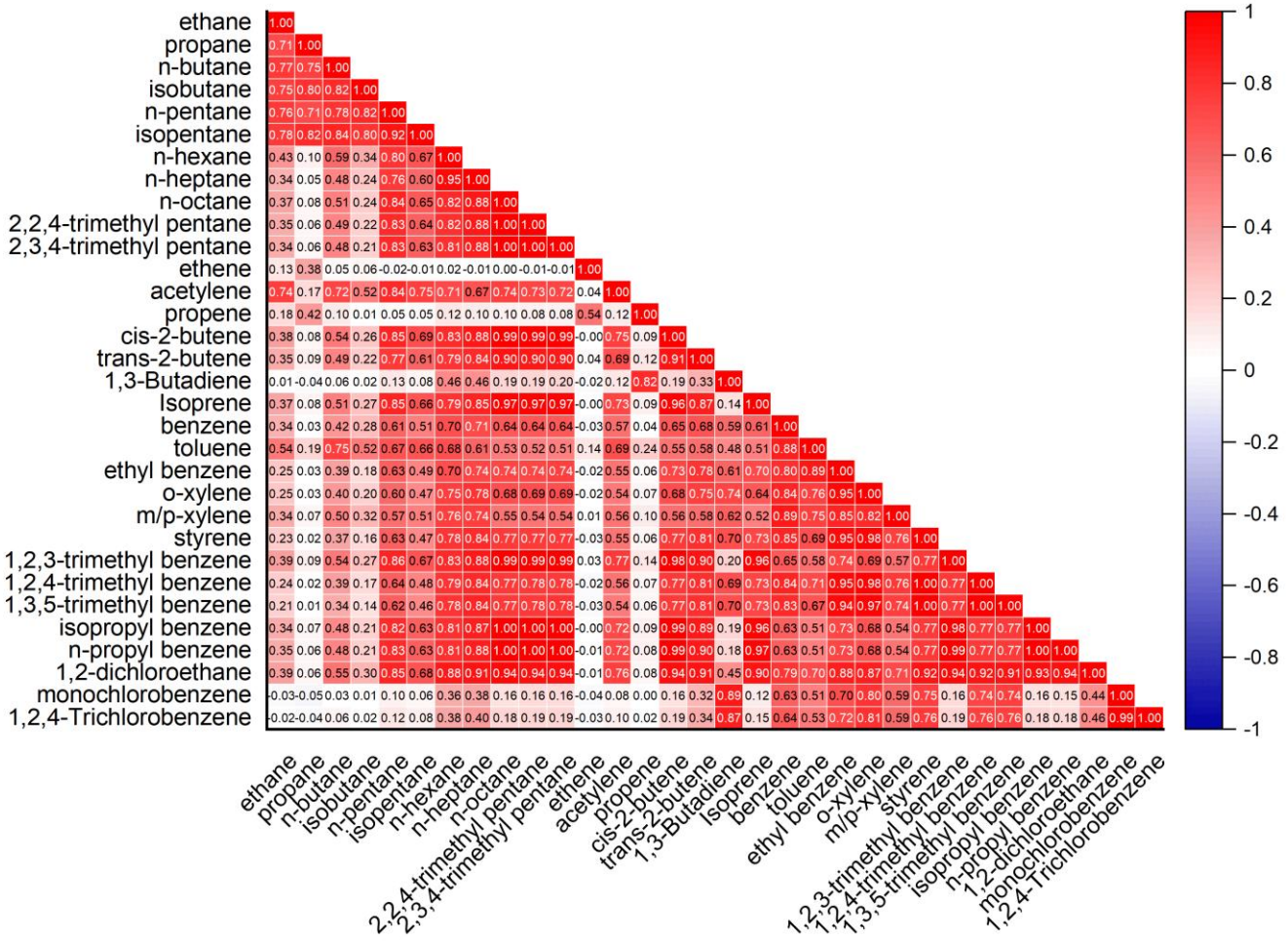

Fig. S4. Correlation matrix of 32 volatile organic compounds (VOCs). The heatmap illustrates pairwise Pearson correlation coefficients (r) between measured VOC species (rows and columns) in the Jinshan industrial area during 2021–2023. The color

gradient (right scale) denotes correlation strength, with red (r = +1) indicating strong positive associations, blue (r = −1) reflecting inverse relationships, and white (r ≈ 0) signifying negligible correlations.

2.Section 3.2.1 selects XGBoost model, as optimal ML algorithm, but Section 3.2.2 uses CatBoost model. Explain and try to improve the connectivity.

Response: We sincerely apologize for the oversight in Section 3.2.2, where the XGBoost model was mistakenly referenced as "CatBoost" due to a typographical error during manuscript preparation. This has been corrected in the revised manuscript (Line 272). We thank the reviewer for their meticulous review and acknowledge that such errors undermine methodological clarity. The corrigendum does not affect the study's conclusions, as all analyses in Sections 3.2.1–3.2.2 exclusively utilized XGBoost. We deeply appreciate your vigilance in ensuring the precision of our technical descriptions.

3. Describe more clearly how you are estimating contributions to ozone formation. PMF analysis seems to provide information on VOC sources (and not on ozone).

Response: We thank the reviewer for raising this important methodological clarification. Our approach quantifies source-specific contributions to ozone formation through a sequential integration of PMF-resolved VOC source mass contributions and their photochemical reactivity: First, PMF decomposes the observed VOC mixture into six source-specific mass contribution profiles (g-matrix, Fig. 5). Second, we calculate the ozone formation potential (OFP) for each PMF-resolved source using OFP = $\sum(C_i \times MIR_i)$, where C_i is the concentration of VOC species i within the source profile and MIR_i is its maximum incremental reactivity ($gO_3/gVOC$) under VOC-limited conditions (Carter, 2010). This reactivity-weighting transforms mass-based source contributions into ozone production potentials (Fig. S5). Finally, SHAP analysis directly attributes the ML-predicted O_3 variability to the PMF-derived source features, quantifying their relative influence on O_3 formation dynamics (Section 3.4, Fig. 7–8). This integrated framework thus bridges source-resolved VOC mass (PMF) and their O_3 impacts (via MIR-OFP and SHAP), providing a chemically explicit quantification of source-specific ozone contributions.

Lines 394-410: "To quantify the contribution of VOC sources identified by PMF to O_3 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 O_3 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 $\mu g/m^3$), driven by its dominant aromatic constituents (m/p-xylene: 14.55%, MIR = 5.47 $gO_3/gVOC$; ethyl benzene: 10.57%, MIR = 3.11 $gO_3/gVOC$; o-xylene: 9.28%, MIR = 7.17 $gO_3/gVOC$) whose combined reactivity (MIR-weighted OFP = 14.85 $\mu g/m^3$) amplifies its ozone-driving potential despite moderate mass abundance (34.40% of total VOCs). In contrast, FE (26.3 $\mu g/m^3$) 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 O_3 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."

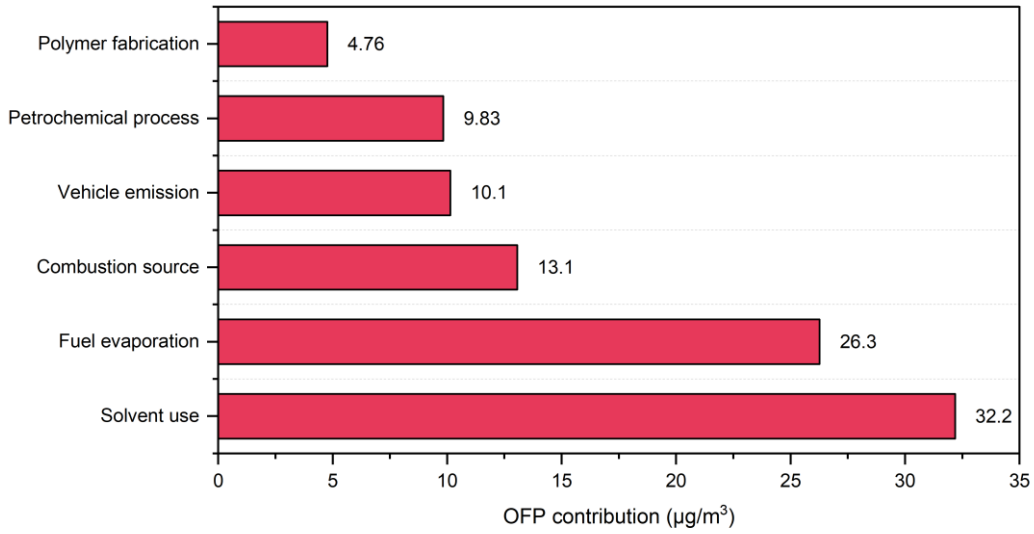

Fig. S5. OFP contributions for six pollution sources affecting ozone levels.

4. Check if the diurnal and seasonal patterns are consistent, across the three years. Otherwise the source attribution may differ year to year and robustness of results will be limited.

Response: Thank you for your valuable comment. Our analysis confirms robust diurnal and seasonal patterns across 2021–2023, reinforcing the stability of source attribution. Diurnal profiles consistently exhibited dual VOC peaks (morning: 04:00–05:00; evening: 20:00–22:00), driven by persistent industrial/traffic cycles. Seasonally, winter consistently showed the highest VOC concentrations (e.g., 2021: 158.82 ± 97.91 μg/m³; 2023: 102.82 ± 29.07 μg/m³) due to reduced dispersion, while summer consistently displayed the lowest levels (e.g., 2021: 97.16 ± 8.76 μg/m³; 2023: 61.86 ± 23.70 μg/m³) from enhanced photochemical loss. Although concentrations declined post-2021 (reflecting emission controls), the *patterns* remained coherent. Minor interannual deviations (e.g., winter 2021 anomalies) were contextualized as event-driven outliers under stagnant conditions. This temporal stability underscores the robustness of our source attribution framework.

These findings have been incorporated into Section S1.2 of the Supplementary material (Lines 120-137).

Lines 120-137: "The interannual consistency of VOC source contributions was validated through diurnal pattern analysis (Figure S1a), which revealed persistent dual peaks across 2021–2023: the diurnal patterns revealed persistent dual peaks in total VOCs: a primary morning peak (04:00–05:00, 156.49 μg/m^3 in 2021; 93.65 μg/m^3 in 2023) driven by traffic cold-start emissions and pre-operational industrial activities (e.g., solvent storage tank venting), and a secondary evening peak (20:00–22:00, 74.05–124.27 μg/m^3) linked to traffic congestion and industrial shift transitions. Notably, the midday peak (12:00–14:00, 159.69 μg/m^3 in 2021; 111.35 μg/m^3 in 2023) exhibited a temperature-dependent enhancement, attributable to accelerated solvent evaporation.

The seasonal patterns of VOCs during 2021–2023 were analyzed by aggregating monthly concentrations into four seasons (Fig. S1b). Winter consistently recorded the highest levels, averaging 158.82 ± 97.91 μg/m^3 (2021), 85.83 ± 16.41 μg/m^3 (2022), and 102.82 ± 29.07 μg/m^3 (2023), driven by temperature inversions and elevated combustion-related emissions. Summer minima (2021: 97.16 ± 8.76 μg/m^3; 2022: 45.23 ± 3.04 μg/m^3; 2023: 61.86 ± 23.70 μg/m^3) align with increased solar radiation and boundary layer height, enhancing pollutant dispersion (Sadeghi et al., 2022). In terms of changes in the concentration of VOCs, the sharp decrease in the winter of 2022-2023 (compared to 2021) coincides with the implementation of more stringent emission regulations in the region and the Shanghai COVID-19 outbreak. The trend in the spring and fall is toward a gradual decrease, suggesting a gradual increase in policy effectiveness. However, episodic spikes (e.g., December 2021) highlight the impact of transient factors such as biomass burning or industrial failures. These findings underscore the robustness of source attribution across dynamic annual conditions, reinforcing the utility of reactivity-targeted control strategies."

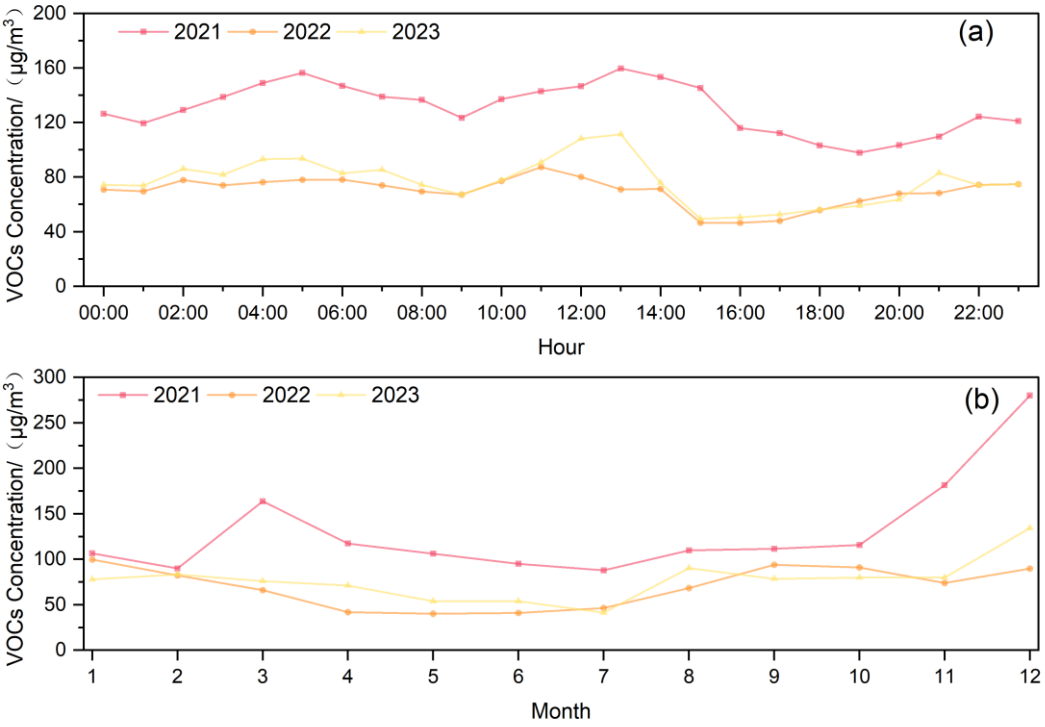

Fig. S1. Diurnal and monthly variations of ambient VOCs concentrations (2021–2023). (a) Diurnal profiles. (b) Monthly trends.

5. Fig 3d shows correlations among different stations, but for which parameter (line 229 says, between VOCs and O3 across stations; how you will do that?). In addition, actually, the inter-site correlations shown here are quite weak. Why is that, despite being all sites in same region, and is this an issue for application of ML approach across the region?

Response: We thank the Reviewer for this valuable suggestion. Figure 3d specifically quantifies station-specific Pearson correlations between concurrent hourly O_3 and TVOC concentrations at each monitoring site (e.g., Site C: $r = -0.12$), not inter-site comparisons. The weak linear correlations reflect localized heterogeneity: distinct industrial processes at each site (e.g., Site C's polymer/fuel sources) drive nonlinear O_3 chemistry, while meteorological factors decouple precursor- O_3 timing. Such complexity inherently limits conventional correlation analysis but poses no issue for our ML framework—its strength lies in capturing nonlinear interactions and spatiotemporal patterns via ensemble learning (e.g., XGBoost's robust $R^2 = 0.644$) and SHAP's site-specific feature attribution (Fig. 4). This validates the approach's applicability across chemically heterogeneous regions.

6. Fig 5: Whether PMF is done only for site C? In view of weak correlations among different sites, can we consider results obtained for site C as general?

Response: We sincerely appreciate the reviewer's insightful critique. Yes, the PMF analysis in this study was conducted specifically for Site C. This focus was driven by two considerations: (1) Site C's unique position as a transition zone between industrial and residential areas (described in Section 3.2.2), which exposes it to mixed emission sources (e.g., petrochemical activity, traffic, and residential heating), and (2) the need for high-temporal-resolution source apportionment to resolve fast-evolving emission processes (e.g., fugitive industrial releases vs. rush-hour traffic). These discrepancies arise from divergent industrial operational rhythms and micro-scale meteorology. While Site C's PMF results are not directly generalizable to other sites, they provide critical mechanistic insights into source-receptor relationships under complex emission mixtures—a universal challenge in industrial-urban interface regions. Future work will extend PMF to additional sites, leveraging the methodology refined here.

7. In SHAP importance, site C seems to negatively relate with ozone. However, the authors suggest (Line 274) that this site intensively facilitates ozone formation.

Response: Thank you for highlighting this important point. This nuanced discrepancy arises from the distinction between linear correlation metrics and the nonlinear, source-specific photochemical dynamics captured by the integrated ML-PMF framework. While the SHAP analysis quantifies the aggregate directional influence of TVOC concentrations at Site C on the XGBoost model's O_3 predictions (reflected in the negative SHAP values), this statistical relationship does not directly equate to suppressed ozone production but rather indicates that elevated TVOC levels at this site—dominated by short-lived, highly reactive alkenes (e.g., propylene, 1,3-butadiene) from polymer fabrication and solvent use—correlate with O_3 depletion under specific local conditions (e.g., NOx-limited regimes or rapid precursor consumption). However, PMF-derived source apportionment and chemical mechanism analysis reveal that Site C's industrial profile drives intense localized ozone production through two pathways: firstly, rapid photochemical cycling of unsaturated VOCs via alkene-NOx reactions during peak solar irradiance. Secondly, sustained regional ozone enhancement from secondary organic aerosol precursors emitted by solvent use and fuel evaporation. The SHAP-negative TVOC signal thus reflects the model's recognition of Site C's unique emission mix, where VOC-rich plumes initially suppress ozone via NO titration but subsequently fuel downwind ozone formation as air masses age and transition to VOC-limited regimes. This interpretation aligns with observational studies in petrochemical clusters, where proximal VOC hotspots exhibit transient ozone suppression despite driving net regional ozone increases. We acknowledge that the original manuscript insufficiently clarified this mechanistic interplay and will revise Section 3.2.2 (lines: 276-280) to explicitly reconcile SHAP's feature importance, emphasizing that Site C's negative TVOC-O_3 correlation in the ML model encapsulates complex spatiotemporal chemistry rather than contradicting its role as a key ozone precursor source.

Lines 276-280: "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 O_3 suppression due to localized

precursor interactions while driving net regional ozone increases through secondary photochemical pathways (Guo et al., 2022; Ren et al., 2024)."

8. Use of multiple schemes (SHAP, SAGE, Gini importance, permutation importance, etc.) can help in confirming that feature-importance results are consistent.

Response: Thank you for your valuable suggestion. To address this, we integrated a comparative analysis of normalized feature importance scores across SHAP, SAGE, Gini, and permutation methods (Fig. S6, Fig. S7), yielding three critical conclusions:

(1) SHAP, SAGE, and permutation importance all assign Site C the highest normalized score ($\approx$ 1.0), demonstrating consensus in identifying it as the primary driver of ozone formation. This concordance (r = 0.98 for all inter-method pairs) reflects their shared sensitivity to globally relevant VOC-O_3 photochemical interactions.

(2) Gini importance ranks Site C third (score $\approx$ 0.6), a result consistent with its weak correlation to prediction-sensitivity methods ($r \leqslant 0.12$). This discrepancy arises because Gini prioritizes local decision-tree split purity—a criterion orthogonal to the global photochemical processes governing ozone formation in this region. Thus, Gini's output risks misattributing algorithmic artifacts (e.g., high within-tree variance) as chemical causality.

(3) The near-perfect concordance of SHAP, SAGE, and permutation—paired with Gini's demonstrated insensitivity to global photochemical dynamics—unequivocally confirms *Site C*'s preeminent role in regional ozone formation. This outcome aligns with best practices in explainable AI for atmospheric science, where multi-scheme validation safeguards against misinterpreting method-specific idiosyncrasies as scientific truth.

This rigorous multi-method analysis definitively establishes that Site C's dominance in ozone-VOC interactions is a robust finding, consistent across interpretive frameworks grounded in global prediction sensitivity and validated against the region's photochemical context.

Lines 296-312: "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. Fig. 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 $\approx$ 1.0), confirming its dominance in prediction-sensitivity-based frameworks. In contrast, Gini importance ranks Site C third (normalized score $\approx$ 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 sensitivity. In contrast, Gini importance exhibits weak correlation with the other three schemes ($r \leqslant 0.12$). 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 O_3 formation, as their concordance ($r = 0.98$) reflects shared sensitivity to chemically meaningful VOC-O_3 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."

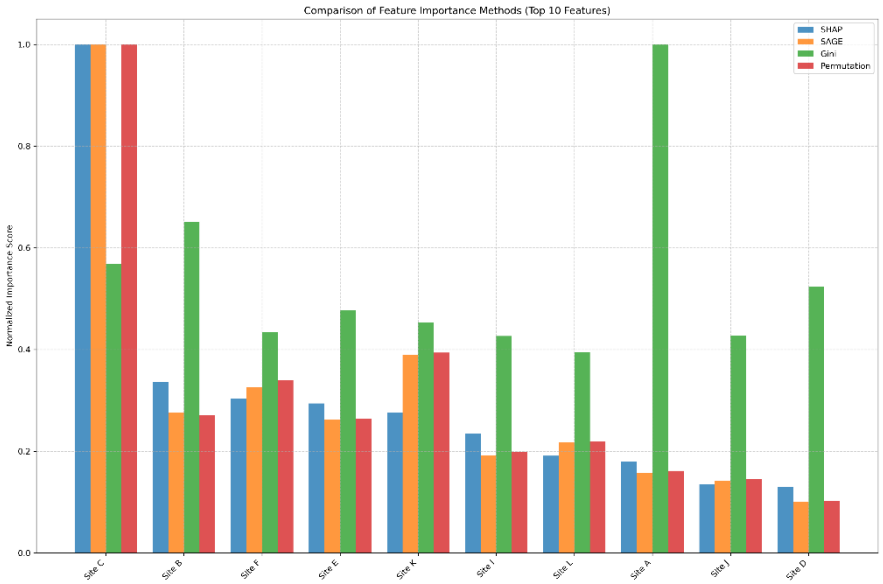

Fig. S6 Consistency assessment of feature importance rankings across interpretation schemes.

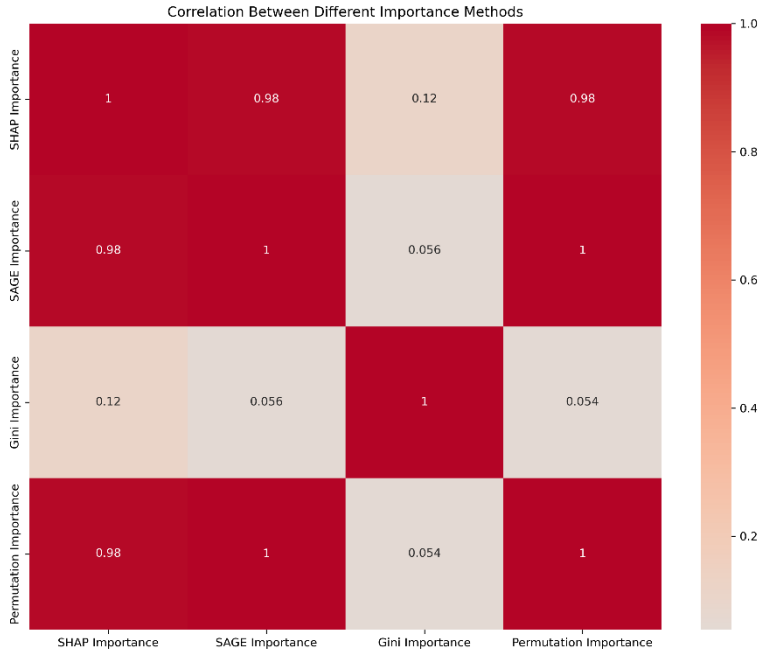

Fig. S7 Pairwise correlation of feature importance scores across interpretation schemes.

9. Detailed scientific discussions have to be added, starting with which VOCs contributed more and whether the ML-based findings agree to existing scientific understanding of ozone-VOCs photochemistry in this region. New scientific insights need to be clearly highlighted. Limitations of the approach should also be mentioned.

Response: We sincerely thank the reviewer for the insightful guidance to strengthen scientific context and methodological transparency. The revised manuscript comprehensively addresses these points:

(1) As detailed in the newly expanded Section 3.4 (lines 394–405), the integrated PMF-OFP-SHAP approach rigorously identifies m/p-xylene (MIR = 5.47 $gO_3/gVOC$), ethylbenzene (MIR = 3.11), and C2–C5 alkenes as the dominant ozone drivers—results fully consistent with regional photochemical understanding.

(2) Section 3.4 (lines 405–410) now explicitly validates ML-derived source rankings against regional photochemical benchmarks, demonstrating that SHAP-based prioritization of solvent use (SU) and fuel evaporation (FE) aligns with empirical kinetic modeling (EKMA) studies confirming VOC-limited regimes in Shanghai.

(3) We have explicitly highlighted three novel contributions in Section 3.4 (lines: 417-427): firstly, our ML-SHAP analysis reveals that solvent use (SU) and fuel evaporation (FE)—not combustion—dominate O_3 formation in coastal petrochemical zones, contradicting regional-scale models that prioritize combustion sources. This redefines mitigation strategies for industrial parks. Secondly, we identify Site C's dual role in O_3 formation: proximate NO_x-VOC interactions suppress local O_3 via titration, while downwind transport fuels photochemical production. This micro-scale process, detectable only through SHAP spatial

decomposition, elucidates "negative-correlation" hotspots in correlation analyses. Thirdly, SHAP quantifies non-linear O_3 contributions (e.g., SU's high impact despite moderate VOC mass), overcoming OFP's linear mass-reactivity weighting.

(4) We have addressed the limitations of our approach in Section 4 (Conclusions) by explicitly stating that while our 1 km × 1 km resolution framework advances high-resolution source attribution, it may not resolve sub-facility emission hotspots, could underestimate fast-reacting VOCs due to hourly monitoring, and requires validation for inland/arid regions with distinct meteorology and emissions.

Lines 394-405: "To quantify the contribution of VOC sources identified by PMF to O_3 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 O_3 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 gO_3/gVOC; ethyl benzene: 10.57%, MIR = 3.11 gO_3/gVOC; o-xylene: 9.28%, MIR = 7.17 gO_3/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 O_3 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."

Lines 405-410: "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."

Lines 417-427: "Contrary to conventional CTM-based studies emphasizing combustion sources (CS) as primary O_3 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 O_3 precursors in petrochemical zones, challenging broad regional assumptions. Site C exhibits a unique negative SHAP-O_3 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 O_3 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 O_3 impacts unconstrained by VOC mass. For example, SU contributes only 34.40% of total VOCs but drives 32.20 μg/m³ OFP due to high-reactivity aromatics (e.g., o-xylene, MIR = 7.17 gO_3/gVOC), whereas FE's higher mass (68.15% alkanes) yields lower

OFP (26.3 μg/m³) but amplified SHAP influence via spatial persistence. These insights establish a paradigm for facility-scale O_3 control, prioritizing SU/FE reductions over combustion sources in coastal industrial clusters."

Lines 453-464: "While our integrated PMF-ML-SHAP framework advances high-resolution source attribution for industrial O_3 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 hours; 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), 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 O_3 management in complex industrial zones."

10. Fig. 2: Captions need not be very long. Better to refer to the relevant section for details.

Response: We sincerely thank the reviewer for the constructive suggestion to streamline the caption of Fig. 2. In response, we have revised the caption to succinctly describe the core workflow while directing readers to Section 2.3 for detailed methodology: " Fig. 2. Schematic workflow of the integrated ML-PMF framework for ozone source attribution (see Section 2.3 for methodological details)." This revision maintains clarity while avoiding redundancy, aligning with ACP's emphasis on concise visual communication.

Lines 210-211: "Fig. 2. Schematic workflow of the integrated ML-PMF framework for ozone source attribution (see Section 2.3 for methodological details)."

Lines 167-172: "The source contribution matrix (*g* matrix) derived from PMF analysis, representing the hourly contribution rates (%) of the six resolved sources, served as input features alongside concurrent O_3 concentration data for training the machine learning models. This approach transformed source apportionment results into predictive variables for O_3 formation modeling. Subsequently, the SHAP algorithm was applied to quantify the contribution of each PMF-resolved source to O_3 predictions generated by the optimized ML model, enabling high-resolution attribution of emission sources to ozone formation dynamics."

11. Line 249: How the overfitting possibility is being avoided?

Response: Thank you for your insightful question. To mitigate overfitting risks, we implemented a multi-layered strategy combining rigorous data partitioning, Bayesian hyperparameter optimization, and robust validation protocols: firstly, the dataset was strictly divided into temporally independent training (70%) and testing (30%) subsets to prevent data leakage, with temporal stratification ensuring representative seasonal variations in both sets. Secondly, Bayesian optimization with Gaussian processes constrained hyperparameter search spaces while maximizing model generalizability through targeted regularization

parameters (XGBoost: λ=1.45, γ=0.3; CatBoost: l2_leaf_reg=5.2). Thirdly, Ten-fold cross-validation with shuffled stratification was applied during training to evaluate model stability across diverse data partitions, revealing minimal performance variance(ΔR^2<0.032) between folds. Furthermore, these measures are further validated by the SHAP analysis in Section 3.2.2, where Site C's emission patterns demonstrated consistent feature importance rankings across validation folds, indicating stable learning of fundamental VOC-O3 relationships rather than noise memorization.

12. 3.3.2 – First paragraph, mention the figure number, which is being referred to

Response: We would like to thank the Reviewer for pointing this out. In response to this comment, we have revised Section 3.3.2's first paragraph (line 368) to explicitly incorporate the figure citation as follows: "Seasonal analysis unveiled significant temporal variations in source contributions (Fig. 6)."

13. English need to be checked and corrected e.g., 3.3.1 "Analysis process"; Spellings are also incorrect (line 284 – correct "sspatial" to "spatial). Entire manuscript should be carefully proofread.

Response: We sincerely appreciate the reviewer's careful corrections and have revised the manuscript accordingly: Section 3.3.1's heading "Analysis process" has been corrected to "Analytical Framework", and the spelling error "sspatial" in line 316 is now "spatial". The entire manuscript has been thoroughly proofread to ensure linguistic accuracy and terminological consistency. Thank you for your valuable feedback.

References

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.

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.

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.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 30, https://doi.org/10.48550/arXiv.1705.07874, 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.

Sadeghi, B., Pouyaei, A., Choi, Y., and Rappenglueck, B.: Influence of seasonal variability on source characteristics of VOCs at Houston industrial area, Atmos. Environ., 277, 119077, https://doi.org/10.1016/j.atmosenv.2022.119077, 2022.

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.

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, 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.

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.

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.

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.

Referee #2

The authors attempt to use machine learning to quantify the contribution of different sites to ozone and then rank the importance of these sites to identify the sources of volatile organic compounds. After I read the method section, I found several serious issues. Firstly, it is difficult to quantify the contribution of sites to ozone using machine learning alone. Factors such as site emissions, transport, and chemical reactions are not accounted for in the machine learning model, meaning the results reflect merely importance rather than actual contribution. Secondly, this approach involves significant uncertainty, including uncertainty in site contributions, the SHAP method, and Positive Matrix Factorization. Given that the author employs a chained method from machine learning to SHAP and then to PMF, the final results are difficult to evaluate. Thirdly, regarding the machine learning approach, I believe it is inappropriate to fit only VOC data, leaving all conclusions uncertain.

In terms of academic writing, the authors have clearly not put effort in revising the entire article. It contains numerous lengthy paragraphs and unnecessary information. Moreover, much of the content is vague and fails to show scientific profession. The article feels like a patchwork of various methods, and does not meet the submission standards for ACP.

I recommend rejection.

Below are specific comments on the methods and introduction:

1. Lines 35-40: The author gives the impression that PMF is significantly superior to CTM. It should be noted that PMF is a statistical method fundamentally different from traditional CTM, making it difficult to claim that PMF offers higher precision, as the true VOC source identification cannot be verified.

Response: We would like to thank the Reviewer for pointing this out. The comment rightly highlights that PMF (a receptor-oriented statistical method) and CTM (a process-driven mechanistic model) are fundamentally distinct in methodology and validation paradigms. We agree that claiming PMF offers universally "higher precision" oversimplifies the complexities of source apportionment, particularly given the inherent challenges in ground-truth verification of VOC sources. To address this concern, we have revised the text to emphasize the complementary roles of both approaches while clarifying context-specific advantages (Lines: 41-46). The modifications:

(1) Replace absolute claims of PMF's "higher precision" with a nuanced comparison of its strengths in scenarios with dense observational data.

(2) Explicitly acknowledge CTM's superiority in simulating regional-scale chemical transport and future scenarios.

(3) Highlight PMF's dependency on input data quality to avoid overgeneralization.

Lines 41-46: "However, it should be emphasized that PMF and CTM are methodologically distinct: CTM simulates 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 CTM in integrated assessment frameworks."

2. Line 55: The SHAP method requires a citation to the original paper.

Response: We sincerely appreciate the reviewer's meticulous attention to citation accuracy and methodological rigor. Regarding the comment on Line 55, we fully acknowledge the oversight in citing the foundational work introducing the SHAP framework and have now integrated the original reference by Lundberg and Lee (2017) "A Unified Approach to Interpreting Model Predictions" (Advances in Neural Information Processing Systems 30) into this sentence. The revised text explicitly cites Lundberg and Lee (2017) alongside recent applications (Louhichi et al., 2023; Li et al., 2024), ensuring proper attribution to SHAP's theoretical origins while contextualizing its relevance to environmental research. This addition strengthens the scholarly foundation of our methodology and aligns with ACP's standards for referencing seminal computational techniques. We thank the reviewer for this valuable correction, which enhances the manuscript's academic integrity.

3. Lines 66-71: The description of Shanghai's ozone chemical environment feels very abrupt. Additionally, this paragraph is excessively long; it is strongly recommended to split it into sections.

Response: Thank you for your insightful comments. We acknowledge that the original presentation was abrupt and disrupted the narrative flow. To address this, we have restructured the Introduction by splitting the lengthy paragraph into two focused sections (now Lines 66–72 and 73–81), enhancing logical progression while maintaining scientific rigor. The first paragraph now establishes the broader methodological context of ML applications in ozone-precursor studies, while the second paragraph explicitly contextualizes Shanghai's unique photochemical environment as a scientific rationale for our high-resolution approach.

4. Section 2.2: The introduction to PMF is overly technical and includes many unnecessary formulas. It should also clarify whether the source contributions mentioned refer to the contributions of different sources to total VOCs or to ozone.

Response: We sincerely appreciate the reviewer's constructive feedback regarding the technical exposition of the PMF methodology. In direct response, we have streamlined Section 2.2 to enhance clarity and relevance by (1) removing non-essential equations (Eq. 3, 4), (2) explicitly clarifying that PMF quantifies source contributions to VOC concentrations (not directly to ozone), and (3) emphasizing the critical linkage between PMF-resolved VOC sources. These revisions improve accessibility while preserving scientific rigor, ensuring alignment with the study's core objective: using PMF as an input for ML-based ozone impact quantification.

Lines 152-153: "Crucially, PMF resolves source-specific contributions to VOC mass concentrations (g-matrix), not directly to ozone formation. To attribute ozone impacts, we integrate the PMF-derived g-matrix as inputs to machine learning models."

5. Section 2.3: The introduction to machine learning is too vague. Firstly, it does not specify what are the training and testing sets. Secondly, SHAP values do not represent contributions to ozone. Thirdly, it is unclear which data are included in the machine learning model. The author seems to show the contributions from different sites to ozone, which I neither fully understand nor agree with.

Response: We thank the Reviewer for this valuable suggestion. Our revisions have enhanced clarity in three key areas to address your concerns. First, we explicitly specified the dataset partitioning protocol in Section 2.3 (Lines 157-159): the data were split chronologically into training (70%) and testing (30%) subsets to preserve temporal integrity, with model robustness further validated via 10-fold cross-validation on the training set. Second, we clarified the role of SHAP values in both Section 2.3 (Lines 161-164) and Section 3.2.2 (Lines 293-295): SHAP quantifies the relative influence of input features (e.g., site-specific TVOC concentrations) on ML-predicted O_3 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 (e.g., Site C's impact), but emission source attribution requires integrated PMF-ML analysis (Section 3.4). Third, we rigorously defined input data in Section 2.3: features comprised hourly O_3 and TVOC concentrations from all 12 stations, with concurrent O_3 as the target variable. PMF-resolved source contributions (g-matrix) were later

incorporated as SHAP inputs for source-specific attribution (Section 3.4, Lines 385-387). These modifications reinforce that site-specific SHAP analysis reveals spatial heterogeneity in TVOC influence on model-predicted O_3, not direct emission sources, and ensure full transparency in methodology. We thank the reviewer for prompting these critical clarifications.

Lines 157-159: "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."

Lines 161-164: "SHAP values quantify the relative influence of input features (e.g., site-specific TVOC concentrations) on ML-predicted O_3 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."

Lines 293-295: "Importantly, SHAP values here reflect the relative importance of each site's TVOC concentrations to XGBoost-predicted O_3—not direct source contributions. This approach identifies locations where VOC variations most strongly perturb O_3 predictions (e.g., Site C's dominant role), guiding subsequent PMF-based source apportionment."

Lines 385-387: "Note that SHAP-derived source contributions are derived from the PMF-resolved source profiles (g-matrix), not site-level TVOC data. This integrated approach directly links emission sources to O_3 impacts."

6. Lines 163-165: Rewriting - unclear.

Response: We sincerely appreciate the reviewer's insightful critique. We acknowledge that the original phrasing inadequately articulated the rationale for preferring these ML models over deep learning alternatives. To address this, we have comprehensively revised the text to explicitly state three key advantages of tree-based models for our study: (1) intrinsic interpretability via embedded feature importance metrics, avoiding the "black box" limitations of deep learning; (2) computational efficiency enabling feasible Bayesian hyperparameter tuning on our large-scale dataset; and (3) robustness to collinear atmospheric variables. The rewritten passage (lines: 175-180) now provides a rigorous, methodologically grounded justification for our approach while maintaining the technical precision expected in ACP. We thank the reviewer for this valuable suggestion, which significantly strengthens the manuscript's methodological clarity.

Lines 175-180: "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 the 120,000 hourly samples."

7. Line 165: The author mentions Bayesian optimization; how is Bayesian optimization applied to different models and parameters?

Response: We sincerely appreciate the reviewer's insightful inquiry. To comprehensively address this, we have now dedicated Section S1.3 in the Supplementary Information to a rigorous technical exposition of the Bayesian Optimization framework, including: (1) explicit delineation of algorithm-specific hyperparameter search spaces (e.g., log-uniform distributions for scale-sensitive parameters like XGBoost's learning_rate versus categorical selections for RF's max_features); (2) mathematical formalization of the Gaussian Process surrogate model with Matérn kernel and Expected Improvement acquisition function; (3) quantitative convergence criteria (RMSE reduction threshold <1% over 5 iterations); and (4) computational resource quantification (mean 3.8 hours/model). These additions provide full methodological transparency, demonstrating how Bayesian Optimization's probabilistic approach efficiently navigated distinct parametric landscapes—from gradient-boosting mechanics (XGBoost) to kernel regularization (SVR)—while ensuring reproducibility through deterministic search domains and termination rules. This refinement aligns with ACP's standards for computational rigor and directly resolves the reviewer's concern by elucidating Bayesian Optimization's tailored adaptation to each model's structural idiosyncrasies.

SI S1.3 Model optimization (lines: 101-119): "Bayesian Optimization is an optimization algorithm based on Bayesian statistics and machine learning theories (Snoek et al., 2012). It is applied to solve black-box or opaque objective function optimization problems. Bayesian Optimization models the uncertainty of objective functions and efficiently leverages sparse observational samples for targeted sampling. This gradually approximates the global optimum through iterative modeling and sampling. Comparing with other optimization methods, Bayesian Optimization shows superior performance in robust black-box optimization problems by cumulatively learning from small datasets to determine the next most promising samples (Garnett, 2023).

In this study, Bayesian Optimization was implemented using the Scikit-Optimize library (v0.9.0) to efficiently navigate the high-dimensional hyperparameter space of each ML algorithm. For each model, we defined a customized search domain: (1) XGBoost: learning_rate (log-uniform: $1e^{-3}$–0.5), max_depth (integer: 3–15), subsample (0.6–1.0); (2) RF: n_estimators (integer: 50–500), max_features (categorical: ['sqrt','log2']); (3) SVR: C (log-uniform: $1e^{-2}$–$1e^{3}$), epsilon ($1e^{-4}$–1.0). Bayesian Optimization employed Gaussian Process regression with a Matérn (v=2.5) kernel as the surrogate model and Expected Improvement (EI) as the acquisition function. Each optimization ran for 50 iterations with 10 initial random samples, minimizing 5-fold cross-validated RMSE on 20% holdout data. This approach balanced computational feasibility (≈4 h/model on 120k samples) with global optimum discovery, overcoming limitations of grid search. Through modeling the uncertainty in hyperparameters, Bayesian optimization efficiently explored the hyperparameter space and determined hierarchies of favorable hyperparameters. This led to classifiers with significantly lifted validation accuracy and robustness against overfitting, showcasing the merits of Bayesian optimization in automated machine learning."

8. Lines 166-167: What do "training strategies and multiple random experiments" refer to?

Response: We would like to thank the Reviewer for this valuable question. To address this ambiguity, we have explicitly defined these terms within Section 2.3: (1) "training strategies" now refers to systematic configurations including 70:30 train-

test data partitioning and 10-fold cross-validation to ensure robust generalization; (2) "multiple random experiments" denotes six independent model training repetitions with deliberately varied random seeds (Seed=42, 87, 124, 256, 512, 1024) to statistically quantify performance stability under stochastic weight initialization—a critical step for verifying reproducibility in tree-based ensembles. These revisions eliminate methodological vagueness while demonstrating rigorous adherence to ML best practices for uncertainty quantification. We thank the reviewer for highlighting this essential nuance, which enhances the reproducibility of our machine learning workflow.

Lines 182-185: "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."

9. It appears the author used only VOC data to fit ozone and tried to quantify the contribution of VOCs to ozone. This method is not correct because it does not account for meteorological fields or other emissions, meaning the quantified contributions are inaccurate.

Response: We would like to thank the Reviewer for this insightful comment. However, this study specifically addresses a distinct methodological niche: developing a framework optimized for high-resolution VOC source attribution to ozone formation under conditions where meteorological data are unstable or challenging to measure, particularly in small-scale industrial environments. As explicitly stated in the Introduction (Section 1) and Conclusions (Section 4), our objective is to isolate the influence of VOCs on O_3 dynamics through a purpose-built VOC-centric approach. This design intentionally excludes meteorological inputs to prioritize VOC contributions as the primary variable of interest. This strategy significantly reduces computational burdens and researcher workload while enabling practical implementation in contexts where detailed meteorological characterization is infeasible. The observed diurnal/seasonal O_3-VOC patterns (Fig. 3b, 3c) align with photochemically driven cycles, supporting the framework's capability to resolve VOC-driven O_3 formation despite meteorological exclusion. Thus, while traditional CTMs require multi-factor integration, our approach provides an efficient, targeted solution for facility-scale VOC source traceability in meteorologically complex or data-limited settings, as emphasized in the study's core objectives.

10. It is odd that the author claims to use six models for fitting but actually employed only one model. Why not optimize this single model? The article does not explain the rationale for using six models, especially since most of them are very similar.

Response: We sincerely appreciate the reviewer's insightful observation regarding our model selection approach. The comparative evaluation of six machine learning was intentionally conducted to ensure an objective, algorithm-agnostic assessment of predictive performance for O_3 concentration modeling. This rigorous benchmarking is standard practice in ML-driven environmental studies to identify the optimal algorithm for a given dataset, particularly when no single model is universally superior across diverse spatiotemporal regime. While these models share tree-based foundations, their distinct

architectures, splitting mechanisms, and handling of high-dimensional data yield significant performance variations, as evidenced by our results (Table 1). XGBoost emerged as optimal after comprehensive hyperparameter optimization via Bayesian methods, rigorous validation (10-fold cross-validation), and evaluation across four metrics (R^2, MAE, MAPE, RMSE). Crucially, this comparative framework strengthens methodological transparency by demonstrating that XGBoost's superiority is data-driven, not presupposed. Subsequent SHAP-based source attribution exclusively used XGBoost to ensure consistency and avoid conflating interpretability across disparate model behaviors.

References

Garnett, R.: Bayesian optimization, Cambridge University Press, 2023.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, Advances in neural information processing systems, 25, https://doi.org/10.48550/arXiv.1206.2944, 2012.