



1 **A Novel Approach Combining Mechanism and Machine**
2 **Learning to Unravel Oxidant-Driven Formation Regimes of**
3 **α -Pinene-Derived Oxygenated Organic Molecules (OOMs)**

4 **Rui Tan^{1,2}, Xinshi Ni¹, Zichao Wan¹, Xudong Tian³, Jiasi Shen³, Bingye Xu³, Da**
5 **Xu³, Sihua Lu¹, Xin Li¹, Song Guo^{1,2*}**

6 ¹ *State Key Laboratory of Regional Environment and Sustainability, International Joint*
7 *Research Center for Atmospheric Research (IJRC), College of Environmental Sciences and*
8 *Engineering, Peking University, Beijing 100871, China*

9 ² *Collaborative Innovation Center of Atmospheric Environment and Equipment Technology,*
10 *Nanjing University of Information Science & Technology, Nanjing 210044, China*

11 ³ *Ecological and Environmental Monitoring Center of Zhejiang Province, Hangzhou 310012,*
12 *China*

13 Correspondence: Song Guo (songguo@pku.edu.cn)

14 **Abstract**

15 Monoterpene oxidation produces oxygenated organic molecules (OOMs) that act as
16 secondary organic aerosol precursors, but the roles of ozone (O₃) and the hydroxyl
17 radical (OH) in their ambient formation are hard to separate because the two oxidants
18 co-vary. We fitted an ensemble of tree-based regression models to one month of online
19 mass-spectrometric measurements at a coastal background site in eastern China and
20 ranked the predictors of four OOM classes defined by hydrogen content (H14–H20)



21 using Shapley additive explanations. Within the fitted models, ozone accounts for 70.8–
22 79.4 % of the relative predictive importance and OH for 20.6–29.2 %. The balance
23 shifts along the series, the class retaining the hydrogen count of the parent monoterpene
24 depending most strongly on ozone and the most saturated class most strongly on OH.
25 The ozone response is not uniform: below about 62 ppb (62.0–63.6 ppb across the four
26 classes) the models extract little ozone signal, whereas above this level ozone becomes
27 a strong positive predictor for all four classes. Three independent checks support this
28 pattern (Appendix B). Both the weights and the transition level derive from a single
29 campaign and are site-specific; the analysis can be repeated wherever these variables
30 are measured.

31 **1 Introduction**

32 The oxidation of volatile organic compounds (VOCs) to low-volatility products
33 governs the formation and evolution of secondary organic aerosol (SOA) (Guo et al.,
34 2020; Wang et al., 2018; Yu et al., 2020; Wang et al., 2021). Among these products,
35 highly oxygenated organic molecules (HOMs), a subclass of oxygenated organic
36 molecules (OOMs), have attracted much attention because of their low volatility and
37 their efficiency in driving new particle formation and growth (Ehn et al., 2014; Kirkby
38 et al., 2016). OOMs are characterized by a high oxygen-to-carbon ratio and by several
39 oxygen-containing functional groups, and can form stable molecular clusters through
40 hydrogen bonding, contributing to homogeneous nucleation (Hallquist et al., 2009;
41 Tang et al., 2025). Monoterpenes are among the most abundant biogenic VOC



emissions worldwide (Sindelarova et al., 2014), and their atmospheric oxidation is initiated mainly by the hydroxyl radical (OH) and ozone (O₃) (Atkinson and Arey, 2003).

Chamber studies have characterized monoterpene oxidation under controlled conditions, but how ambient OOM formation responds to changing oxidant regimes is still poorly constrained. Quantifying this response is complicated by the strong co-variability of oxidants, precursors and boundary-layer dynamics in the ambient atmosphere (Shrivastava et al., 2017; Mcfiggans et al., 2019). The concentrations of OH and O₃ vary strongly in space and time, and their oxidation products are mixed during subsequent chemical processing (Pye et al., 2019). Established approaches such as tracer ratios and box model calculations are limited by uncertainties in the chemical mechanism, by unidentified intermediates and by the limited dimensionality of the observational data (Canonaco et al., 2015; Zhang et al., 2019).

Here we apply a data-driven framework to the oxidant-dependent response of monoterpene-derived OOM formation, together with the accompanying sensitivities to temperature, relative humidity and nitrogen oxides. The analysis uses continuous measurements with a Filter Inlet for Gases and AEROSols coupled to an iodide chemical ionization mass spectrometer (FIGAERO-I-CIMS) during a one-month campaign at a coastal background site in the Yangtze River Delta region of eastern China, with concurrent meteorological and trace gas measurements (Appendix A). An ensemble of tree-based regression models was fitted separately for each of four OOM



63 classes and interpreted with Shapley additive explanations (SHAP), which quantify the
64 relative importance of the predictors and resolve the shape of the ozone response.
65 Because no functional form is prescribed, the models can reveal behavior whose
66 governing chemistry is itself the object of study, and the resulting statistical
67 dependencies are then examined against mechanistic expectations. The input variables
68 are restricted to quantities measured routinely at background and air-quality stations –
69 ozone, nitrogen oxides, temperature, relative humidity, a photolysis-based OH proxy
70 and monoterpene – together with the mass-spectrometric abundances of the product
71 classes, so the same diagnostic can be applied wherever these are available.

72 **2 Results**

73 **2.1 Molecular-class heterogeneity in oxidant sensitivity**

74 A total of 23 monoterpene OOMs in the range $C_{10}H_{14-20}O_{5-10}$ were identified and
75 grouped into four homologous series according to hydrogen content: H14, H16, H18
76 and H20. The diurnal variations of ozone, monoterpene and the total OOM
77 concentrations (gas plus particle phase) are shown in Fig. S2, and representative mass
78 spectra and the mean concentration shares of the four classes in Figs. S3 and S4
79 (Molteni et al., 2019).

80 All four classes are produced mainly during daytime, when OH and O_3 are both
81 abundant. The diurnal amplitude decreases systematically from H14 to H20, which
82 suggests that the more saturated species are less directly coupled to instantaneous
83 photochemical production or have longer atmospheric lifetimes, consistent with



84 formation through several oxidation steps or through ageing. None of the classes shows
85 a distinct secondary night-time maximum, indicating that dark ozonolysis alone does
86 not produce a clear nocturnal maximum at this coastal background site (Zha et al., 2018).
87 This class-to-class heterogeneity, together with the day–night contrast, limits what low-
88 time-resolution bulk measurements and linear correlation analysis can resolve (Peng et
89 al., 2024), and calls for a nonlinear treatment of the OH and O₃ contributions. Because
90 the classes are defined by hydrogen content, an intrinsic property of the products rather
91 than of the site, the same class-resolved analysis can be repeated wherever product
92 abundances and the main drivers are measured together, although the weights
93 themselves will differ from site to site.

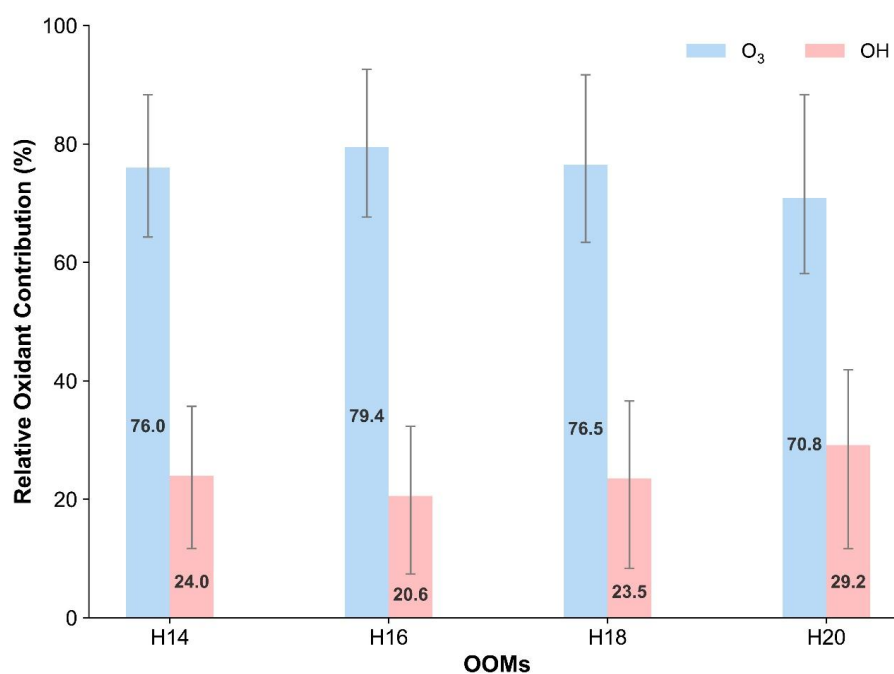
94 **2.2 Quantitative attribution of OH and O₃ oxidation**

95 The ensemble framework and SHAP interpretation described in Appendix A were used
96 to separate the modelled sensitivities of the OH- and O₃-initiated pathways from the
97 influence of other environmental variables; the models serve as an explanatory tool
98 rather than for prediction. SHAP values were averaged over the five independently
99 trained and cross-validated algorithms, and the ensemble mean and standard deviation
100 are reported so that the derived dependencies do not rest on a single algorithm.
101 Temperature, relative humidity, NO and monoterpene were included as control
102 variables, and their contributions are not interpreted here.

103 The relative predictive importance, taken from the ensemble mean of absolute SHAP
104 values (Fig. 1), shows ambient O₃ to be the dominant statistical predictor of OOM



105 abundance at this coastal site, with 70.8–79.4 % of the relative predictive importance
 106 across the four classes, and OH a secondary driver with 20.6–29.2 %. Within the fitted
 107 models, the ozonolysis channel therefore carries about three times the explanatory
 108 weight of the OH channel.



109
 110 Figure 1. Machine learning-derived relative predictive contributions of O₃ and OH to
 111 the formation of different OOMs classes. The error bars represent the model variance
 112 (standard deviation) derived from the five ensemble models.

113 The analysis also resolves systematic differences between the classes. H16, which
 114 retains the hydrogen count of the parent monoterpene (C₁₀H₁₆), shows the strongest
 115 dependence on O₃ (79.4 %) and the weakest on OH (20.6 %): within the model,
 116 products that preserve the carbon skeleton through the initial ozone addition are

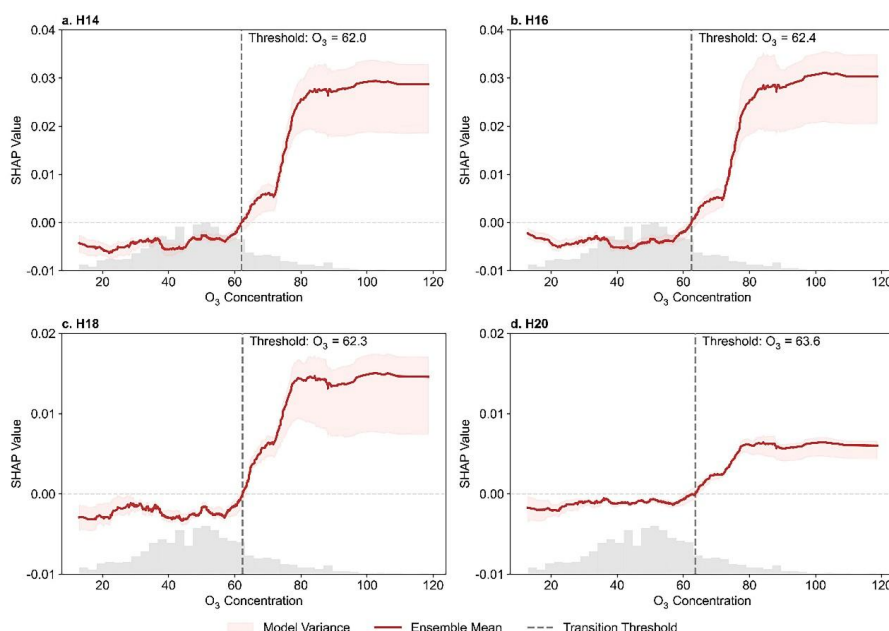


117 associated mainly with the ozonolysis channel. H14 reaches higher daytime
118 concentrations and has a slightly larger diurnal amplitude than H16 (Fig. S2) but a
119 similar oxidant split (76.0 % O₃, 24.0 % OH). The model cannot identify which step
120 initiates the oxidation, so the similarity indicates only that both classes draw most of
121 their explained variance from ozone.

122 The most saturated class, H20, shows the opposite signature: the lowest ozone
123 importance (70.8 %) and the highest OH contribution (29.2 %), consistent with its weak
124 diurnal amplitude (Fig. S2). Adding the hydrogen atoms needed to form C₁₀H₂₀ species
125 requires either several generations of oxidation, which integrates repeated OH exposure
126 over time, or reduction reactions mediated by peroxy radicals (RO₂), so the stronger
127 OH dependence points to more aged and more deeply processed conditions. The use of
128 OOM composition as an indicator of the prevailing chemical regime has been proposed
129 previously; the class-resolved weights obtained here provide a quantitative counterpart
130 to that idea (Piletic and Kleindienst, 2022; Zang et al., 2023; Zhang et al., 2024).

131 **2.3 Nonlinear response of OOM formation to ozone**

132 SHAP dependence analysis (Fig. 2) resolves how the modelled ozone contribution
133 varies across the ozone range instead of averaging over it, as a linear regression does.
134 The transition point is defined as the ozone mixing ratio at which the ensemble-mean
135 SHAP dependence curve crosses zero from negative to positive values, with linear
136 interpolation between the two adjacent points bracketing zero; it is an approximate
137 statistical transition point rather than a chemical threshold.



138

139 Figure 2. SHAP dependence analysis of O_3 concentration for H14, H16, H18, and H20

140 OOMs classes. The solid red line indicates the ensemble mean SHAP value across the

141 five ensemble models. The shaded pink area represents the model variance, calculated

142 as the standard deviation of the SHAP values across the five ensemble models. The grey

143 vertical dashed line denotes the statistical transition point defined in the text. The grey

144 histogram at the bottom of each panel shows the data frequency distribution of the

145 observed O_3 concentrations. Units: O_3 concentration is in ppb.

146 Below this level the SHAP values are close to zero or slightly negative, so ozone

147 variations contribute little to the explained variance of OOM abundance; above it they

148 increase steeply and ozone becomes a primary explanatory factor for all four classes.

149 The four transition levels lie within 1.6 ppb of one another (62.0–63.6 ppb), so the data

150 do not resolve any ordering among the classes. The measured abundances behave



151 consistently with this pattern: mean total OOM mass is about $0.11 \mu\text{g m}^{-3}$ below 62 ppb
152 and about $0.18 \mu\text{g m}^{-3}$ above 64 ppb (Fig. S8, Appendix B).

153 **3 Discussion**

154 A plausible reading of this pattern is that the amount of low-volatility product formed
155 per unit of monoterpene lost to ozonolysis changes with the prevailing oxidant regime.
156 At low ozone (below about 50 ppb), which at this site is usually accompanied by higher
157 NO, termination of peroxy radicals through $\text{RO}_2 + \text{NO}$ can suppress the autoxidation
158 chain that builds up the oxygenated products. Ozonolysis still proceeds, but the yield
159 of OOMs from that channel stays low, so ozone variability carries little information
160 about the observed product abundance. Above about 62 ppb, where NO is generally
161 lower, the termination channel is less competitive, the autoxidation pathway is less
162 suppressed and ozone becomes a sensitive predictor within the model. The
163 measurements do not resolve the RO_2 chemistry directly, so this interpretation is
164 consistent with the statistical pattern but is not established by it.

165 The campaign-mean ozone mixing ratio was 50.1 ± 17.1 ppb, so the site spent much of
166 the record below the 62–63 ppb transition range and, within the fitted models, ozone
167 carries little explanatory weight over a large part of that range. If the same behavior
168 holds elsewhere, the conversion of monoterpene emissions into low-volatility products
169 would vary with the prevailing oxidant regime rather than being fixed for a given
170 landscape. That matters for regions where background ozone is rising: if the transition
171 is a general feature, the same monoterpene emissions would be converted into low-



172 volatility products more efficiently as ozone increases. Repeating the analysis at sites
173 with different ozone and NO_x backgrounds is the direct test of this suggestion.

174 Chamber experiments have quantified α -pinene ozonolysis yields under controlled
175 oxidant levels (Ehn et al., 2014; Kirkby et al., 2016; Molteni et al., 2019) and have
176 shown that single-precursor experiments differ from multi-component mixtures
177 (McFiggans et al., 2019). Those experiments vary one oxidant at a time and report near-
178 steady-state yields, whereas the present analysis estimates the share of the modelled
179 explanatory weight carried by the ozonolysis channel, about three quarters (Fig. 1), with
180 O₃, OH and NO varying together. Two related developments are the use of highly
181 oxygenated product composition as an indicator of ozone formation sensitivity (Zhang
182 et al., 2024) and the finding that class-resolved yields depend on the branching between
183 radical termination and autoxidation (Zang et al., 2023). Both are bulk or linear
184 descriptors, while the weights reported here are resolved per homologous series and are
185 nonlinear, so the two kinds of quantity address different questions.

186 Three checks constrain the attribution from different directions: the analysis repeated
187 separately for the gas and particle phases, an independent kinetic estimate of the
188 ozonolysis branching ratio, and the measured abundance contrast across the transition.
189 They are reported in Appendix B; none is decisive on its own.

190 **4 Conclusions**



191 Ozone is the dominant statistical predictor of monoterpene-derived OOM abundance at
192 this coastal background site, carrying 70.8–79.4 % of the relative predictive importance
193 in the fitted models, with OH accounting for 20.6–29.2 %. The two oxidants are not
194 interchangeable: the class that retains the hydrogen count of the parent monoterpene
195 depends most strongly on ozone, whereas the most saturated class draws the largest
196 share of its explained variance from OH. The ozone response is nonlinear, with the
197 modelled ozone signal concentrated above about 62 ppb and little ozone contribution
198 below it. These weights are properties of the fitted models and the transition level is an
199 empirical result from a single campaign, so neither should be carried into regional or
200 global parameterizations without testing elsewhere. The result provides an observation-
201 based benchmark for the relative roles of the two oxidants under ambient conditions,
202 against which explicit gas-phase mechanisms can be examined, for instance with
203 observationally constrained box models.

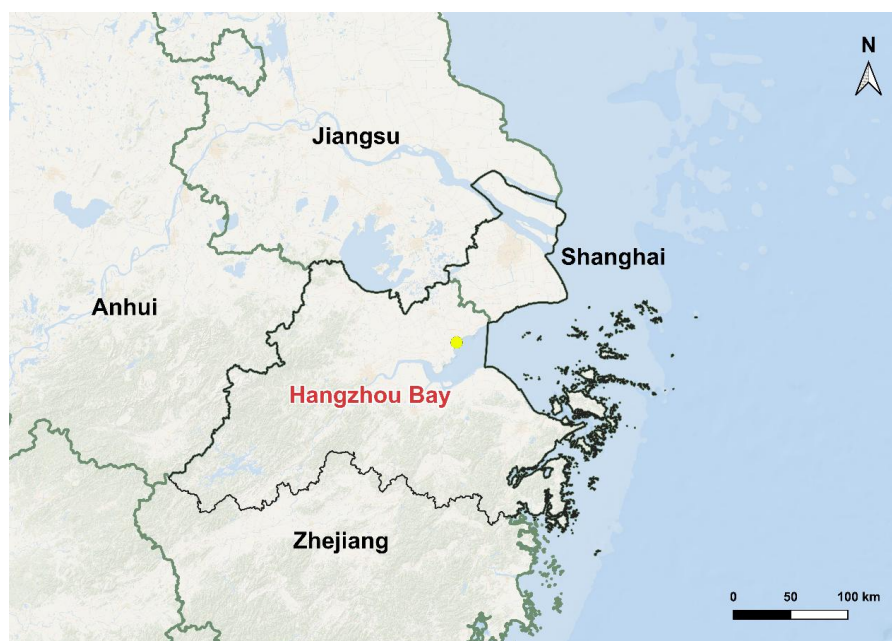
204 **Appendix A: Measurement and modelling approach**

205 **A1 Site and measurements**

206 Field measurements were conducted from 20 April to 28 May 2024 at a regional
207 background site in Haiyan, Zhejiang Province, in the Yangtze River Delta, China
208 (120°59'49" E, 30°33'40" N; Fig. SA1). The site represents the background conditions
209 of the Hangzhou Bay area and is largely removed from immediate local anthropogenic
210 emissions. The campaign was scheduled for late spring, when biogenic activity and
211 photochemical processing are both active (Hu et al., 2026).



212 The site's coastal location, with the East China Sea to the east and southeast, induces
 213 periodic sea-land breeze circulations that significantly govern local meteorology and
 214 regional air mass transport. Complementing this coastal influence, the surrounding land
 215 cover is dominated by mixed forest with a high density of coniferous species,
 216 particularly pine trees. These forests serve as a prolific local source of α -pinene, a key
 217 monoterpene. Collectively, these geographical and ecological features render the site
 218 well suited for investigating biogenic VOC (BVOC) oxidation chemistry under
 219 representative background conditions, characterized by relatively limited local
 220 anthropogenic emissions compared to urban centers within the Yangtze River Delta.



221

222

Figure A1. Location of the Haiyan site.

223 Meteorological parameters (temperature, relative humidity, pressure, wind speed and



224 wind direction) were monitored continuously. Nitrogen oxides (NO, NO₂ and NO_x),
225 ozone (O₃) and PM_{2.5} mass concentrations were measured with standard commercial
226 analyzers.

227 Ambient volatile organic compounds (VOCs), including monoterpenes, were measured
228 with a Vocus proton transfer reaction time-of-flight mass spectrometer (VOCUS-PTR-
229 ToF, Tofwerk AG / Aerodyne Research Inc., hereafter VOCUS). Oxygenated organic
230 molecules in both the gas and the particle phase were measured with a Filter Inlet for
231 Gases and AEROSols coupled to an iodide-adduct chemical ionization mass
232 spectrometer (FIGAERO-I-CIMS, Tofwerk AG / Aerodyne Research Inc.). The
233 VOCUS reports the sum of the monoterpene isomers that share the elemental
234 composition C₁₀H₁₆, so the oxidation products discussed below are described as
235 monoterpene derived, which reflects this instrumental limitation rather than a speciated
236 precursor measurement.

237 The FIGAERO-I-CIMS enables simultaneous measurement of gas-phase OOMs and
238 particle-phase composition via thermal desorption of collected particles. The maximum
239 thermal desorption temperature reached around 180 °C, ensuring that most volatile and
240 semi-volatile species could be effectively desorbed and analyzed. Based on the
241 characteristics of iodide-adducted chemical ionization, the detected species are
242 predominantly compounds containing functional groups such as hydroxyl and nitro
243 groups, which exhibit high affinity for iodide ions. The OOMs species detected in this
244 study were predominantly in the range of C₁₀H₁₄₋₂₀O₅₋₁₀. Species with higher degrees
245 of oxygenation beyond this range were not reliably quantified using iodide-adducted
246 chemical ionization.



247 The FIGAERO-I-CIMS operated with a time resolution of 1 hour. During each hourly
248 cycle, the first 20 minutes were dedicated to gas-phase sampling, while particles were
249 simultaneously collected onto a quartz filter at a flow rate of approximately 3 L min⁻¹.
250 The subsequent 40 minutes were allocated to particle-phase analysis, which consisted
251 of the following steps: (1) a 1-minute switching and stabilization period to ensure stable
252 operation after mode transition; (2) a 14-minute ramping period during which the filter
253 temperature was increased linearly from room temperature to approximately 180 °C;
254 (3) a 13-minute constant heating period at 180°C to ensure complete thermal desorption
255 of all collected species; and (4) a 12-minute cooling period to return the filter to ambient
256 temperature for the next sampling cycle.

257 For the quantification of both gas-phase and particle-phase species, a unified sensitivity
258 was applied. In this study, the sensitivity of levoglucosan was calibrated using a
259 micro-syringe injection method. Levoglucosan has been widely reported in the
260 literature as a compound approaching the maximum sensitivity achievable with I-CIMS.
261 Therefore, we adopted the levoglucosan sensitivity as the uniform sensitivity for all
262 detected OOMs species. This approach yields a lower-limit estimate of the absolute
263 concentrations of the target species.

264 **A2 Machine learning framework**

265 To quantify the nonlinear dependencies of OOM formation on the environmental
266 drivers, an interpretable data-driven framework was used. Mechanism-based methods
267 such as tracer ratios or box models require a prescribed reaction scheme and rate
268 coefficients and are therefore less suited to detecting behavior whose governing
269 chemistry is itself under investigation; a supervised learning model makes no
270 assumption about the functional form linking drivers to products, and its interpretation
271 can be tested against mechanistic expectations. Separate models were built for each



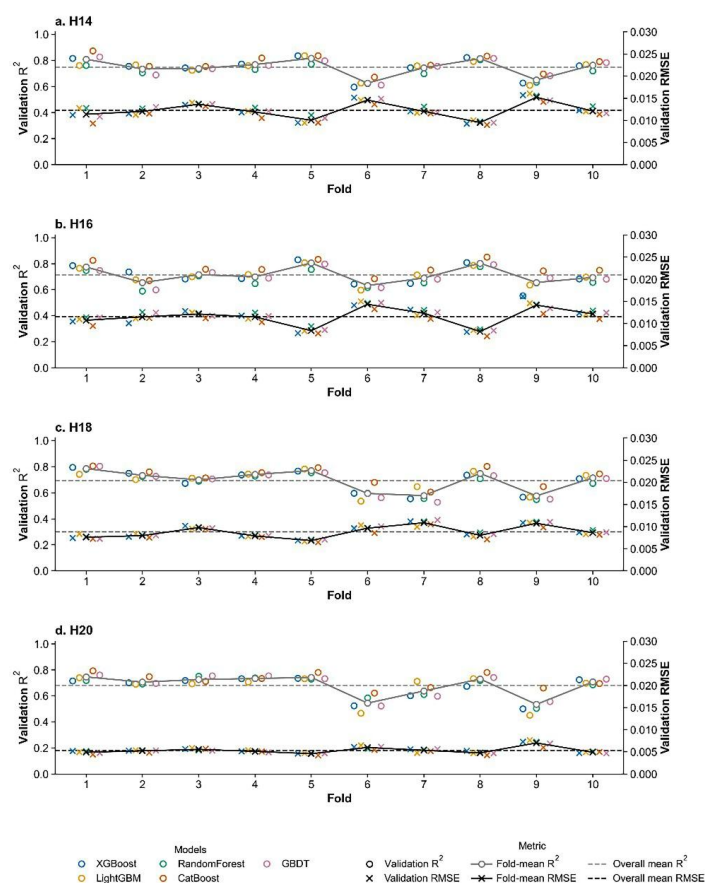
272 homologous series of monoterpene oxidation products, namely $C_{10}H_{14}O_{\geq 5}$, $C_{10}H_{16}O_{\geq 5}$,
273 $C_{10}H_{18}O_{\geq 5}$ and $C_{10}H_{20}O_{\geq 5}$ (H14–H20), with the sum of the gas-phase and particle-phase
274 concentrations of each class as the target variable.

275 An ensemble of five structurally different tree-based regression algorithms, spanning
276 the boosting and bagging families, was used to reduce the dependence of the result on
277 any single algorithm; the individual algorithms are listed below. The input features were
278 the ozone concentration (O_3), a hydroxyl radical proxy ($J(O^1D)/[NO_2]$, where $J(O^1D)$ is
279 the photolysis rate constant of O_3 to $O(^1D)$ and $[NO_2]$ is the ambient NO_2 concentration;
280 the proxy is evaluated in Sect. A3), the monoterpene concentration, temperature,
281 relative humidity (RH) and nitric oxide (NO). Hyperparameters were tuned for each
282 class by Bayesian optimization combined with 10-fold cross-validation, maximizing
283 the coefficient of determination (R^2); fold-wise validation results are shown in Fig. SA2.

284 Five advanced tree-based ensemble learning algorithms—namely eXtreme Gradient
285 Boosting (XGBoost), Categorical Boosting (CatBoost), Light Gradient Boosting
286 Machine (LightGBM), Random Forest (RF), and Gradient Boosting Decision Tree
287 (GBDT)—were deployed (Chen and Guestrin, 2016; Prokhorenkova et al., 2018; Gong
288 et al., 2025; Ileri, 2025; Yu et al., 2025; Belgiu and Dragut, 2016; Khajavi and Rastgoo,
289 2023; Luo et al., 2025). Although these algorithms share a tree-based structure, they
290 differ fundamentally in their optimization frameworks. XGBoost, CatBoost,
291 LightGBM, and GBDT belong to the Boosting family that iteratively reduces bias,
292 whereas Random Forest is a Bagging method that reduces variance. By aggregating the
293 SHAP values from the ensemble of these five structurally diverse models, we
294 effectively eliminate algorithmic bias and enhance the robustness of the feature
295 importance rankings.



For each OOM class, we adopted a 10-fold cross-validation framework, dividing the dataset into 10 subsets. In each iteration, 9 folds were used for model fitting, and the remaining 1 fold, which was not involved in the fitting process, was used for performance evaluation. The R^2 and RMSE values reported are calculated from these held-out validation folds.



301

302 Figure A2. Validation performance (R^2 and RMSE) of the five machine learning
303 models across 10-fold cross-validation for the H14, H16, H18, and H20 OOM classes.

304 The left y-axis shows the Validation R^2 , and the right y-axis shows the Validation
305 RMSE. In each subplot, the colored open circles represent the R^2 of individual models
306 on each validation fold, and the colored crosses represent the RMSE of individual



307 models. The solid black lines indicate the fold-mean R^2 and RMSE across the five
308 ensemble models. The horizontal dashed lines denote the overall mean R^2 and RMSE
309 across all 10 folds and all five models. The model colors correspond to XGBoost (blue),
310 Random Forest (green), CatBoost (orange), LightGBM (yellow), and GBDT (pink)

311 Figure A2 shows the fold-wise validation results for H14, H16, H18, and H20 across
312 the five machine learning algorithms. In the figure, the colored symbols represent the
313 R^2 and RMSE of different models on each validation fold; the solid colored lines
314 represent the average validation performance of the five models per fold; and the
315 horizontal dashed lines denote the overall average validation performance across all
316 folds and all models. The results show that the validation R^2 for all OOM classes
317 remains positive and consistently high across the 10 folds, while the fold-to-fold
318 variation in RMSE is relatively limited. No systematic collapse in validation
319 performance was observed. Furthermore, the five different algorithms exhibit
320 consistent variation patterns, indicating that, within this ensemble, the identified
321 statistical relationships do not depend on a specific model architecture.

322 To ensure the reliability of our SHAP attributions, potential multicollinearity among the
323 input features was quantitatively examined by calculating the Variance Inflation Factor
324 (VIF) for all predictors. The results showed all VIF values to be well below the critical
325 threshold of 5, confirming that multicollinearity does not significantly affect the feature
326 importance estimates derived from the tree-based ensemble models.

327 After the best model for each target variable had been selected, SHAP (SHapley
328 Additive exPlanations), a model-agnostic interpretability method, was applied to
329 quantify the relative contribution of each feature. SHAP values quantify the marginal
330 statistical contribution of a feature to the modelled OOM abundance and do not imply
331 chemical causation, so the resulting dependencies are treated as statistical explanatory
332 importance within the model. SHAP dependence analysis was then used to characterize



the nonlinear response patterns, and these statistical dependencies were checked against an independent kinetic estimate and against the measurements (Appendix B). A variance inflation factor analysis indicated no significant multicollinearity among the input features.

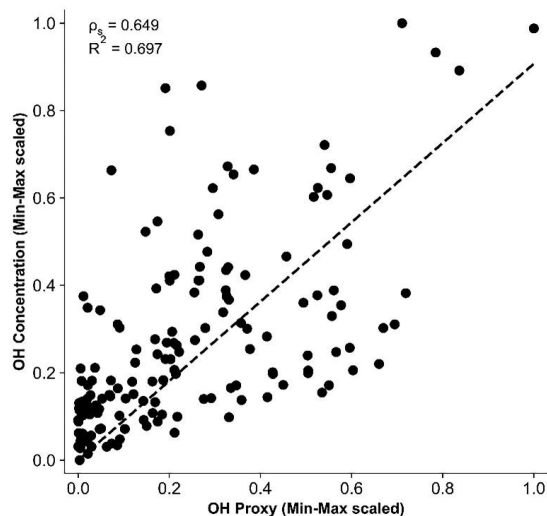
A3 Evaluation of the hydroxyl radical proxy

Direct measurements of OH radical concentrations were not available for the entire campaign. Therefore, a photochemical proxy was constructed based on the steady-state relationship between OH production and loss. Previous field observations have demonstrated a strong positive correlation between OH radical concentrations and the photolysis rate constant of O₃ to O(¹D), across various environments (Tan et al., 2017; Berresheim et al., 2003; Lu et al., 2012). This correlation is known to vary with NO_x levels. To account for the influence of NO_x on OH sinks, we adopt the ratio $J(O^1D)J(O^1D)/NO_2$ as a OH proxy.

$$[OH]_{proxy} = \frac{J(O^1D)}{[NO_2]}$$

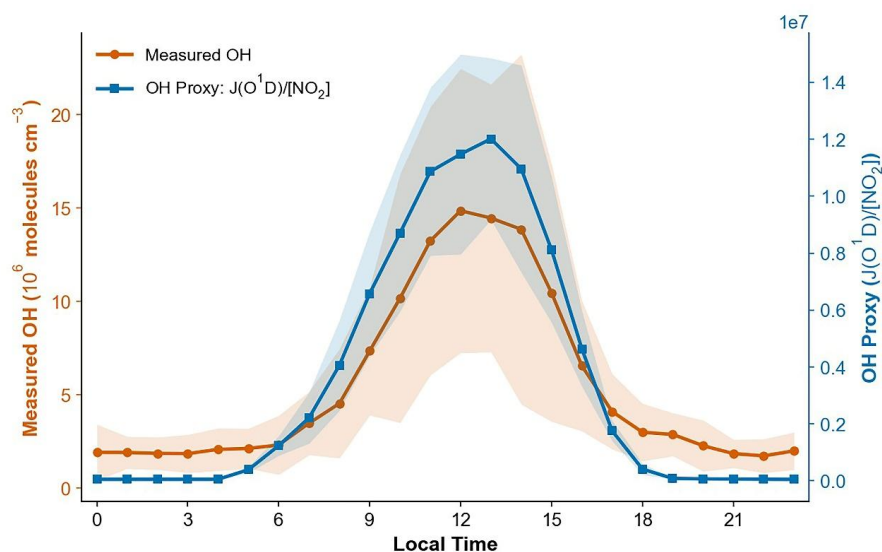
Where $[OH]_{proxy}$ is the proxy parameter of OH concentration; $J(O^1D)$ is the photolysis rate constant of O₃ to O¹D, measured by a J-VALUE spectrometer; $[NO_2]$ is the concentration of NO₂. This proxy captures the diurnal variation of OH driven by solar radiation (via directly measured $J(O^1D)$) and modulated by NO₂ levels, without requiring explicit knowledge of the absolute OH loss rate.

Figure A3 presents the comparison between the $[OH]_{proxy}$ and the LIF-measured OH concentrations during the final week. Spearman correlation $\rho_s=0.649$, $R^2=0.697$, indicating that the proxy captures OH variation well.



355

356 Figure A3. Comparison of OH proxy indicators and OH concentration.



357

358 Figure A4. Diurnal variation of the OH proxy ($J(O^1D)/[NO_2]$) compared with the
 359 directly measured OH concentration by Laser-Induced Fluorescence (LIF). The left y-
 360 axis shows the measured OH concentration ($10^6 \text{ molecules cm}^{-3}$), and the right y-axis



361 shows the OH proxy value. The shaded areas represent the variation ranges.

362 **Appendix B: Independent checks on the model-derived oxidant** 363 **weights**

364 As a first check, the analysis was repeated separately for the gas- and particle-phase
365 fractions of each class (Fig. S7, Sect. S2). The ordering of the two oxidants is preserved
366 in both phases: within the models, O₃ accounts for 77.6–79.6 % of the relative
367 predictive importance in the gas phase and 62.6–69.6 % in the particle phase, while the
368 OH proxy rises from 20.4–22.4 % to 30.4–37.4 %. The direction of the shift is what the
369 multi-generation interpretation in Sect. 2.2 predicts: products that have undergone more
370 oxidative processing, and that therefore partition more readily into the particle phase,
371 draw a larger share of their modelled variance from the OH channel. Because the two
372 phases come from the same campaign rather than from independent measurements, this
373 check constrains the internal consistency of the attribution rather than confirming it
374 against separate data.

375 As a second check, the SHAP weights were compared with a kinetic estimate that does
376 not involve the model (Sect. S3). The mean kinetic fraction of α -pinene lost to O₃ during
377 the campaign was 53.2 %, obtained from ambient oxidant concentrations and the
378 bimolecular rate coefficients. Weighting this branching ratio by pathway-specific molar
379 yields gives a theoretical O₃ contribution of about 89.7 %, an upper bound because it
380 neglects other oxidants such as NO₃ and assumes no radical sinks. The SHAP-derived
381 range (70.8–79.4 %) falls below that bound, which is what would be expected if



382 ambient loss processes, particularly $\text{RO}_2 + \text{NO}$ termination, reduce the effective
383 contribution of the ozonolysis pathway. The two estimates differ in timescale and in
384 what they measure, so the comparison is qualitative: it shows that the statistical
385 dominance of O_3 is of a magnitude the independent kinetics can accommodate, not that
386 the two are numerically equivalent.

387 A third check asks whether the transition identified in Sect. 2.3 is visible in the
388 measurements themselves (Fig. S8, Sect. S4). Mean total OOM abundance is about 0.11
389 $\mu\text{g m}^{-3}$ in the low- O_3 regime (< 62 ppb, $n = 687$) and about 0.18 $\mu\text{g m}^{-3}$ in the high- O_3
390 regime (> 64 ppb, $n = 156$). The separation is in the direction the SHAP dependence
391 curves imply, but the two regimes differ in sample size and in their NO_x and temperature
392 distributions, and the comparison of the two means carries no formal test; it therefore
393 supports the transition rather than establishing it.

394 **Code availability**

395 The code used for the machine-learning analysis is available from the corresponding
396 author upon reasonable request.

397 **Data availability**

398 The data needed to reproduce Figs. 1 and 2 are provided in the accompanying data file
399 submitted together with this manuscript.

400 **Author contributions**

401 Conceptualization: R.T. Data curation: R.T. and Z.W. Formal analysis: R.T. and X.N.



Investigation: R.T., X.N. and Z.W. Methodology: R.T. and X.N. Software: X.N. and R.T.
Writing – original draft: R.T. Writing – review and editing: X.T., J.S., B.X., D.X., S.L. X.L.
and S.G. Supervision: S.G. Funding acquisition: S.G. All authors have read and agreed to the
published version of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors thank the staff of the Ecological and Environmental Monitoring Center of
Zhejiang Province for their support during the Haiyan field campaign.

Financial support

This work was supported by the National Key R&D Program of China
(2022YFC3701002, 2022YFC3703502), National Natural Science Foundation of
China-Creative Research Group Fund (22221004), special fund of State Key
Laboratory of Regional Environment and Sustainability (Grant No. 26Y01RESPKU).

References

- Atkinson, R. and Arey, J.: Atmospheric Degradation of Volatile Organic Compounds,
Chemical Reviews, 103, 4605-4638, 10.1021/cr0206420, 2003.
- Belgiu, M. and Dragut, L.: Random forest in remote sensing: A review of applications and
future directions, Isprs Journal of Photogrammetry and Remote Sensing, 114, 24-31,
10.1016/j.isprsjprs.2016.01.011, 2016.
- Berresheim, H., Plass-Dülmer, C., Elste, T., Mihalopoulos, N., and Rohrer, F.: OH in the
coastal boundary layer of Crete during MINOS: Measurements and relationship with
ozone photolysis, Atmos. Chem. Phys., 3, 639-649, 10.5194/acp-3-639-2003, 2003.
- Canonaco, F., Slowik, J. G., Baltensperger, U., and Prévôt, A. S. H.: Seasonal differences in
oxygenated organic aerosol composition: implications for emissions sources and factor
analysis, Atmos. Chem. Phys., 15, 6993-7002, 10.5194/acp-15-6993-2015, 2015.
- Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proceedings of the
22nd ACM SIGKDD International Conference on Knowledge Discovery and Data



- 430 Mining, San Francisco, California, USA, 10.1145/2939672.2939785, 2016.
- 431 Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M.,
432 Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen,
433 M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T.,
434 Kurtén, T., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M.
435 D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R.,
436 Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol,
437 Nature, 506, 476-479, 10.1038/nature13032, 2014.
- 438 Gong, J. Q., Qu, Z. G., Zhu, Z. L., Xu, H. T., and Yang, Q. G.: Ensemble models of TCN-
439 LSTM-LightGBM based on ensemble learning methods for short-term electrical load
440 forecasting, Energy, 318, 10.1016/j.energy.2025.134757, 2025.
- 441 Guo, S., Hu, M., Peng, J., Wu, Z., Zamora, M. L., Shang, D., Du, Z., Zheng, J., Fang, X.,
442 Tang, R., Wu, Y., Zeng, L., Shuai, S., Zhang, W., Wang, Y., Ji, Y., Li, Y., Zhang, A. L.,
443 Wang, W., Zhang, F., Zhao, J., Gong, X., Wang, C., Molina, M. J., and Zhang, R.:
444 Remarkable nucleation and growth of ultrafine particles from vehicular exhaust,
445 Proceedings of the National Academy of Sciences, 117, 3427-3432,
446 10.1073/pnas.1916366117, 2020.
- 447 Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M.,
448 Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann,
449 H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A.,
450 Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H.,
451 Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of
452 secondary organic aerosol: current and emerging issues, Atmos. Chem. Phys., 9, 5155-
453 5236, 10.5194/acp-9-5155-2009, 2009.
- 454 Hu, K., Guo, S., Shen, L., Wan, Z., Ni, X., Tan, R., Tian, X., Xu, B., Zeng, L., Fu, Q., Hu, M.,
455 and Lu, S.: Underestimation of industrial contributions to VOC fluctuation in the
456 Hangzhou Bay: Insights from simultaneous measurements and ML-SHAP analysis,
457 Environmental Pollution, 400, 128282, <https://doi.org/10.1016/j.envpol.2026.128282>,
458 2026.
- 459 Ileri, K.: Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods
460 optimised with PSO to estimate the number of k-barriers for intrusion detection in
461 wireless sensor networks, International Journal of Machine Learning and Cybernetics, 16,
462 6937-6956, 10.1007/s13042-025-02654-5, 2025.
- 463 Khajavi, H. and Rastgoo, A.: Predicting the carbon dioxide emission caused by road transport
464 using a Random Forest (RF) model combined by Meta-Heuristic Algorithms, Sustainable
465 Cities and Society, 93, 10.1016/j.scs.2023.104503, 2023.
- 466 Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., Heinritzi, M.,
467 Simon, M., Yan, C., Almeida, J., Troestl, J., Nieminen, T., Ortega, I. K., Wagner, R.,
468 Adamov, A., Amorim, A., Bernhammer, A.-K., Bianchi, F., Breitenlechner, M., Brilke, S.,
469 Chen, X., Craven, J., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida,
470 R., Hakala, J., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Kim, J., Krapf,
471 M., Kuerten, A., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Molteni, U.,
472 Onnela, A., Peraekylae, O., Piel, F., Petaejae, T., Praplan, A. P., Pringle, K., Rap, A.,



- 473 Richards, N. A. D., Riipinen, I., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger,
474 S., Scott, C. E., Seinfeld, J. H., Sipilä, M., Steiner, G., Stozhkov, Y., Stratmann, F.,
475 Tome, A., Virtanen, A., Vogel, A. L., Wagner, A. C., Wagner, P. E., Weingartner, E.,
476 Wimmer, D., Winkler, P. M., Ye, P., Zhang, X., Hansel, A., Dommen, J., Donahue, N. M.,
477 Worsnop, D. R., Baltensperger, U., Kulmala, M., Carslaw, K. S., and Curtius, J.: Ion-
478 induced nucleation of pure biogenic particles, *Nature*, 533, 521–, 10.1038/nature17953,
479 2016.
- 480 Lu, K. D., Rohrer, F., Holland, F., Fuchs, H., Bohn, B., Brauers, T., Chang, C. C., Häseler, R.,
481 Hu, M., Kita, K., Kondo, Y., Li, X., Lou, S. R., Nehr, S., Shao, M., Zeng, L. M., Wahner,
482 A., Zhang, Y. H., and Hofzumahaus, A.: Observation and modelling of OH and
483 HO₂ concentrations in the Pearl River Delta 2006: a missing OH source in a
484 VOC rich atmosphere, *Atmos. Chem. Phys.*, 12, 1541–1569, 10.5194/acp-12-1541-2012,
485 2012.
- 486 Luo, J. Q., Yuan, Y., and Xu, S. X.: Improving GBDT performance on imbalanced datasets:
487 An empirical study of class-balanced loss functions, *Neurocomputing*, 634,
488 10.1016/j.neucom.2025.129896, 2025.
- 489 McFiggans, G., Mentel, T. F., Wildt, J., Pullinen, I., Kang, S., Kleist, E., Schmitt, S., Springer,
490 M., Tillmann, R., Wu, C., Zhao, D., Hallquist, M., Faxon, C., Le Breton, M., Hallquist,
491 Å. M., Simpson, D., Bergström, R., Jenkin, M. E., Ehn, M., Thornton, J. A., Alfarra, M.
492 R., Bannan, T. J., Percival, C. J., Priestley, M., Topping, D., and Kiendler-Scharr, A.:
493 Secondary organic aerosol reduced by mixture of atmospheric vapours, *Nature*, 565, 587–
494 593, 10.1038/s41586-018-0871-y, 2019.
- 495 Molteni, U., Simon, M., Heinritzi, M., Hoyle, C. R., Bernhammer, A.-K., Bianchi, F.,
496 Breitenlechner, M., Brilke, S., Dias, A., Duplissy, J., Frege, C., Gordon, H., Heyn, C.,
497 Jokinen, T., Kürten, A., Lehtipalo, K., Makhmutov, V., Petäjä, T., Pieber, S. M., Praplan,
498 A. P., Schobesberger, S., Steiner, G., Stozhkov, Y., Tomé, A., Tröstl, J., Wagner, A. C.,
499 Wagner, R., Williamson, C., Yan, C., Baltensperger, U., Curtius, J., Donahue, N. M.,
500 Hansel, A., Kirkby, J., Kulmala, M., Worsnop, D. R., and Dommen, J.: Formation of
501 Highly Oxygenated Organic Molecules from α -Pinene Ozonolysis: Chemical
502 Characteristics, Mechanism, and Kinetic Model Development, *ACS Earth and Space*
503 *Chemistry*, 3, 873–883, 10.1021/acsearthspacechem.9b00035, 2019.
- 504 Peng, Z., Zhang, B., Wang, D., Niu, X., Sun, J., Xu, H., Cao, J., and Shen, Z.: Application of
505 machine learning in atmospheric pollution research: A state-of-art review, *Science of The*
506 *Total Environment*, 910, 168588, <https://doi.org/10.1016/j.scitotenv.2023.168588>, 2024.
- 507 Piletic, I. R. and Kleindienst, T. E.: Rates and Yields of Unimolecular Reactions Producing
508 Highly Oxidized Peroxy Radicals in the OH-Induced Autoxidation of α -Pinene, β -Pinene,
509 and Limonene, *The Journal of Physical Chemistry A*, 126, 88–100,
510 10.1021/acs.jpca.1c07961, 2022.
- 511 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A.: CatBoost:
512 unbiased boosting with categorical features, 32nd Conference on Neural Information
513 Processing Systems (NIPS), Montréal, Canada, Dec 02–08, 2018.
- 514 Pye, H. O. T., D'Ambro, E. L., Lee, B. H., Schobesberger, S., Takeuchi, M., Zhao, Y., Lopez-
515 Hilfiker, F., Liu, J., Shilling, J. E., Xing, J., Mathur, R., Middlebrook, A. M., Liao, J.,



- 516 Welti, A., Graus, M., Warneke, C., de Gouw, J. A., Holloway, J. S., Ryerson, T. B.,
517 Pollack, I. B., and Thornton, J. A.: Anthropogenic enhancements to production of highly
518 oxygenated molecules from autoxidation, *Proceedings of the National Academy of*
519 *Sciences*, 116, 6641-6646, doi:10.1073/pnas.1810774116, 2019.
- 520 Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L.,
521 Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J.,
522 Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang,
523 J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in
524 understanding secondary organic aerosol: Implications for global climate forcing,
525 *Reviews of Geophysics*, 55, 509-559, <https://doi.org/10.1002/2016RG000540>, 2017.
- 526 Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrou, T., Müller, J. F.,
527 Kuhn, U., Stefani, P., and Knorr, W.: Global data set of biogenic VOC emissions
528 calculated by the MEGAN model over the last 30 years, *Atmos. Chem. Phys.*, 14, 9317-
529 9341, 10.5194/acp-14-9317-2014, 2014.
- 530 Tan, Z., Fuchs, H., Lu, K., Hofzumahaus, A., Bohn, B., Broch, S., Dong, H., Gomm, S.,
531 Häseler, R., He, L., Holland, F., Li, X., Liu, Y., Lu, S., Rohrer, F., Shao, M., Wang, B.,
532 Wang, M., Wu, Y., Zeng, L., Zhang, Y., Wahner, A., and Zhang, Y.: Radical chemistry at a
533 rural site (Wangdu) in the North China Plain: observation and model calculations of OH,
534 HO₂ and RO₂ radicals, *Atmos. Chem. Phys.*, 17, 663-690, 10.5194/acp-17-663-2017,
535 2017.
- 536 Tang, L., Feng, Z., Shang, D., Zeng, L., Wu, Z., Wang, H., Chen, S., Li, X., Zeng, L., Hu, J.,
537 and Hu, M.: Ongoing uncoordinated anthropogenic emission abatement promotes
538 atmospheric new particle growth in a Chinese megacity, *Nature Communications*, 16,
539 6720, 10.1038/s41467-025-62011-6, 2025.
- 540 Wang, H., Xiang, Z., Wang, L., Jing, S., Lou, S., Tao, S., Liu, J., Yu, M., Li, L., Lin, L., Chen,
541 Y., Wiedensohler, A., and Chen, C.: Emissions of volatile organic compounds (VOCs)
542 from cooking and their speciation: A case study for Shanghai with implications for China,
543 *Sci Total Environ*, 621, 1300-1309, 10.1016/j.scitotenv.2017.10.098, 2018.
- 544 Wang, H., Guo, S., Yu, Y., Shen, R., Zhu, W., Tang, R., Tan, R., Liu, K., Song, K., Zhang, W.,
545 Zhang, Z., Shuai, S., Xu, H., Zheng, J., Chen, S., Li, S., Zeng, L., and Wu, Z.: Secondary
546 aerosol formation from a Chinese gasoline vehicle: Impacts of fuel (E10, gasoline) and
547 driving conditions (idling, cruising), *Science of The Total Environment*, 795, 148809,
548 <https://doi.org/10.1016/j.scitotenv.2021.148809>, 2021.
- 549 Yu, T. K., Chang, I. C., Chen, S. D., Chen, H. L., and Yu, T. Y.: Predicting potential soil and
550 groundwater contamination risks from gas stations using three machine learning models
551 (XGBoost, LightGBM, and Random Forest), *Process Safety and Environmental*
552 *Protection*, 199, 10.1016/j.psep.2025.107249, 2025.
- 553 Yu, Y., Wang, H., Wang, T., Song, K., Tan, T., Wan, Z., Gao, Y., Dong, H., Chen, S., Zeng, L.,
554 Hu, M., Wang, H., Lou, S., Zhu, W., and Guo, S.: Elucidating the importance of semi-
555 volatile organic compounds to secondary organic aerosol formation at a regional site
556 during the EXPLORE-YRD campaign, *Atmospheric Environment*, 118043,
557 <https://doi.org/10.1016/j.atmosenv.2020.118043>, 2020.
- 558 Zang, H., Huang, D., Zhong, J., Li, Z., Li, C., Xiao, H., and Zhao, Y.: Direct probing of



559 acylperoxy radicals during ozonolysis of α -pinene: constraints on radical chemistry and
560 production of highly oxygenated organic molecules, *Atmos. Chem. Phys.*, 23, 12691-
561 12705, 10.5194/acp-23-12691-2023, 2023.

562 Zha, Q., Yan, C., Junninen, H., Riva, M., Sarnela, N., Aalto, J., Quéléver, L., Schallhart, S.,
563 Dada, L., Heikkinen, L., Peräkylä, O., Zou, J., Rose, C., Wang, Y., Mammarella, I., Katul,
564 G., Vesala, T., Worsnop, D. R., Kulmala, M., Petäjä, T., Bianchi, F., and Ehn, M.: Vertical
565 characterization of highly oxygenated molecules (HOMs) below and above a boreal
566 forest canopy, *Atmos. Chem. Phys.*, 18, 17437-17450, 10.5194/acp-18-17437-2018,
567 2018.

568 Zhang, J., Zhao, J., Luo, Y., Mickwitz, V., Worsnop, D., and Ehn, M.: On the potential use of
569 highly oxygenated organic molecules (HOMs) as indicators for ozone formation
570 sensitivity, *Atmos. Chem. Phys.*, 24, 2885-2911, 10.5194/acp-24-2885-2024, 2024.

571 Zhang, Y., Peräkylä, O., Yan, C., Heikkinen, L., Äijälä, M., Daellenbach, K. R., Zha, Q., Riva,
572 M., Garmash, O., Junninen, H., Paatero, P., Worsnop, D., and Ehn, M.: A novel approach
573 for simple statistical analysis of high-resolution mass spectra, *Atmos. Meas. Tech.*, 12,
574 3761-3776, 10.5194/amt-12-3761-2019, 2019.

575