Supplement of Atmos. Chem. Phys., 25, 9601–9615, 2025
https://doi.org/10.5194/acp-25-9601-2025-supplement

Atmospheric Chemistry and Physics
Open Access

Supplement of

Influencing factors of the gas–particle distribution of oxygenated organic molecules in the urban atmosphere and deviation from equilibrium partitioning: a random forest model study

Xinyu Wang et al.

Correspondence to: Huan Yu (yuhuan@cug.edu.cn)

Table of Contents

Text S1 Routine measurement of gaseous and particulate components

In this study, we measured the chemical composition of $PM_{2.5,}$ including water-soluble ions (sulfate (SO_4^{2-}), nitrate (NO_3^-), ammonium (NH_4^+), chloride (Cl^-), and potassium ion (K^+)) over 594 hours using an Online Ion Chromatography Monitoring System (MARGA-1S, Metrohm AG, Switzerland) for the water-soluble ions. The system is designed to collect and analyze $PM_{2.5}$ in ambient air in real-time. Air samples are first passed through a cutter and sampling tubes into the instrument, where aerosol particles are captured and mixed with water vapor in a high-temperature steam generator. This process causes the particles to grow by absorption, after which they condense and are directed to the sample collection unit. The collected samples are then separated and analyzed using ion chromatography for their water-soluble ion content, including both cations and anions.

The carbonaceous materials were analyzed using an OCEC Analyzer (RT-4, Sunset Laboratory Inc., USA), which employs a stepwise heating pyrolysis-oxidation method. The sample is first heated under helium (He) gas, causing the OC to volatilize and partially convert to pyrolyzed carbon (PC). The sample is then further heated in a helium/oxygen (He/O_2) mixture, where EC is oxidized and decomposed into gaseous oxidation products. All decomposition products flow through a carrier gas into an oxidation furnace, where the carbon products are converted to CO_2 and quantitatively detected using non-dispersive infrared (NDIR) methods. During this process, laser transmittance is used to monitor the OC/EC separation point, with OC volatilization and pyrolysis causing a decrease in transmittance intensity and EC oxidation leading to an increase. When the transmittance intensity returns to its initial level, the OC/EC separation point is defined, allowing for the precise determination of OC and EC content in the sample.

$PM_{2.5}$ concentrations were measured using an air particulate monitor (TH-2000PM, Wuhan Tianhong Technology Co., Ltd.), utilizing a dual-channel beta-ray method coupled with a dynamic compensation system. The air flow, set to 33.34 L/min, is first passed through a PM_{10}-cutting device to separate the larger particles. The flow is then evenly split into two streams: one stream directly measures PM_{10}, while the other is passed through a $PM_{2.5}$-cutting device for the measurement of $PM_{2.5}$. This approach reduces measurement errors and ensures accurate quantification of both PM_{10} and $PM_{2.5}$ concentrations.

Hourly concentrations of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), and ammonia (NH_3) were detected using gas analyzers (42i/43i/49i/17i, Thermo Fisher Scientific, USA). SO_2 was measured via pulsed fluorescence technology, in which SO_2 molecules absorb ultraviolet light at a specific wavelength and re-emit fluorescent light. The intensity of the emitted fluorescence is directly proportional to the SO_2 concentration. O_3 was quantified using ultraviolet photometry, as ozone absorbs ultraviolet light at a specific wavelength (254 nm). The O_3 concentration is determined by measuring the intensity of the absorbed light. For NH_3 measurement, the method involves its reaction with oxygen at high temperatures (750°C), converting NH_3 into nitrogen monoxide (NO). The NO concentration is directly proportional to the NH_3 concentration, which is then calculated accordingly. The concentration of NO_2 is measured using chemiluminescence: the sample first passes through a molybdenum catalyst at 325°C to convert NO2 into NO, which then reacts with O_3 generated by a silent discharge ozone generator, producing chemiluminescence detected by a photomultiplier tube (PMT). By measuring in NO and NO_x modes—recording the NO concentration without catalysis and the total NO (including converted NO from NO_2) with catalysis—the NO_2 concentration is calculated as the difference between the two measurements.

Meteorological parameters, including relative humidity (RH), temperature (T), wind speed (WS), and wind direction (WD), were collected using an automatic weather station. Photolysis frequencies of HONO (JHONO) were measured with a PFS-100 photolysis spectrometer (Focused Photonics Inc., China). The spectrometer uses a quartz receiver head to collect solar radiation from various directions and transmits the light through optical quartz fibers to the spectrometer. The spectrometer then transmits the spectral data to an industrial computer, which calculates the photolysis flux. By integrating this flux with known absorption cross-sections and quantum yields, the photolysis rate constant is determined.

Text S2 Determination of thermal decomposition fragments

Some studies, like Stark et al. (2017), suggested that a fraction of oxygenated organic molecules (OOMs) detected by Chemical Ionization Mass Spectrometer equipped with a Filter Inlet for Gases and AEROsol (FIGAERO-CIMS) may be due to thermal decomposition. To explore this issue, we employed a K-means algorithm (Faxon et al., 2018) , which is an unsupervised machine learning approach, to cluster the OOMs to 5 groups High-Volatility Low-Mw OOMs, and Low-Volatility Low-Mw OOMs, High-Volatility Median-Mw OOMs, Low-Volatility Median-Mw OOMs, High-Mw OOMs and decomposition fragment on the basis of thermal desorption temperature T_{max}, molecular weight (Mw), carbon number (n_C) and concentration measured at the urban supersite during the winter campaign. Here the concentration of an OOM is the average concentration during the campaign (Wang et al., 2024). The input variables of the K-means algorithm were scaled to values between 0 and 1, based on their respective numerical ranges, to prevent any bias associated with the relative magnitude of each variable. The clustering results can be seen in Figure S7.

Text S3 Parameterizations of saturated mass concentration C^*

The C^* parameterizations from Mohr et al. (2019) and Ren et al. (2022) are given by Equation (S1):

$$\log_{10} C^* = (n_C^0 - n_C)b_C - (n_O - 3n_N)b_O - 2 \times \frac{(n_O - 3n_N)n_C}{(n_C + n_O - 3n_N)} b_{CO} - n_N b_N \qquad \text{(S1)}$$

where n_C^0 is the reference carbon number; n_C, n_O, and n_N are the numbers of carbon, oxygen, and nitrogen atoms in an organic species, respectively; b_C, b_O, and b_N denote the contribution of each kind of atoms to $\log_{10} C^*$, respectively, and b_{CO} represents the carbon-oxygen nonideality.
the carbon–oxygen nonideality, The coefficients from the studies of Mohr et al. and Ren et al. are shown in Table S6.

The parameterizations from Peräkylä et al. (2020) and Priestley et al. (2024) are given by Equations S2 and S3, respectively:

$$\log_{10} C^* = 0.18 \times n_C - 0.14 \times n_H - 0.38 \times n_O + 0.8 \times n_N + 3.1 \qquad \text{(S2)}$$

$$\log_{10} C^* = -0.03 \times n_C - 0.08 \times n_H - 0.14 \times n_O - 0.14 \times n_N + 4.26 \qquad \text{(S3)}$$

Text S4 Calculation of aerosol pH and liquid water content (LWC)

To calculate pH and LWC, the ISORROPIA-II model in forward mode was used in this study (Fountoukis and Nenes, 2007). Input parameters included RH (0-1 scale), temperature (K), and concentrations (mol m^{-3}) of K^+, Ca^{2+}, Mg^{2+}, total ammonia (NH_4^+ in the particulate phase and NH_3 in the gas phase), total sulfate (SO_4^{2-} in the particulate), total sodium (Na^+), total chloride (Cl^- in the particulate phase and HCl in the gas phase), and total nitrate (NO_3^- in the particulate phase and HNO_3 in the gas phase).

The model calculates aerosol pH using Equation S4 (Guo et al., 2016):

$$pH = -log_{10}(\frac{[H^+]_{air} \times \gamma_{H^+}}{LWC_{mol\,m^{-3}} \times 0.018}) \qquad \text{(S4)}$$

where γ_{H^+} is the hydronium ion activity coefficient (assumed = 1), $[H^+]_{air}$ represents the concentration of free H^+ ions per unit air volume (mol m^{-3}), and $LWC_{mol\,m^{-3}}$ is the aerosol liquid water content (mol m^{-3}). The 0.018 corresponds to the volume occupied by 1 mol of water, with units of L mol^{-1}. Both $[H^+]_{air}$ and $LWC_{mol\,m^{-3}}$ are obtained from thermodynamic model calculations.

Text S5 Outlier removal

The criterion for outlier removal is based on the Interquartile Range (IQR) rule, which involves the following steps: (1) Calculate the interquartile range for the data, (2) Multiply the interquartile range by 1.5, (3) Subtract 1.5 × the interquartile range from the first quartile—any value less than this threshold is considered an outlier and is removed, and (4) Add 1.5 × the interquartile range to the third quartile—any value greater than this threshold is considered an outlier and is removed from the original data.

Text S6 SHAP analysis

Complex machine learning models, such as the random forest model, often achieve higher predictive accuracy compared to simpler models like linear models, but this comes at the cost of reduced interpretability. Properly interpreting the outputs of these complex models is crucial, as it allows for better application and trust in the model's predictions (Ribeiro et al., 2016).

SHAP is a unified framework for post hoc model interpretation based on cooperative game theory (Lundberg, 2017; Jia et al., 2024). The core idea of SHAP is to calculate the marginal contribution of features to the model's output, thereby explaining the "black box" model. SHAP can interpret model outputs at both the global and local levels, meaning it can explain the average importance of each feature across the entire model as well as the specific impact of each feature on individual predictions . For each prediction sample i, the output of the random forest model can be expressed as:

$$y_i = y_{base} + \sum_{j=1}^{k} f(x_{ij}) \qquad \text{(S5)}$$

Where:$y_i$is the predicted value for sample i, y_{base} is the baseline value for the model, typically the

mean prediction across all samples, k represents the number of features used by the model, x_{ij} is the j-th feature for sample i, $f(x_{ij})$ is the SHAP value of feature x_{ij}, which represents its marginal contribution to the prediction.

SHAP values are calculated by evaluating the marginal contribution of each feature to the model's output. In the SHAP framework, each feature's contribution can be positive or negative, depending on whether it increases or decreases the prediction. SHAP values not only quantify the importance of features but also reflect the positivity or negativity of the influence of features.

We used the SHAP method to quantify the importance of influential features and to calculate the marginal effects of a single feature on the Gas-to-Particle (G/P) ratios of OOMs. SHAP values measure the marginal contribution of a feature to the model's prediction, considering the contributions of all other features. A positive SHAP value indicates a feature's positive contribution to enhancing the prediction, while a negative value indicates a negative contribution. The magnitude of the SHAP value indicates the feature's impact on the prediction. SHAP results aid in assessing the model's decision-making process, identifying important features, and discovering underlying data patterns. The SHAP values were computed using the shap package (v.0.40.0) in Python (v.3.8). The data were then fitted with a Generalized Additive Model (GAM) using the pygam package (v.0.8.0).

Text S7 Partial Dependence Plots

The partial dependence plots (PDP) is a method for interpreting black-box models (Shi et al., 2023). Mathematically, the partial dependence function for the j-th feature is defined as:

$$\widehat{f}_j(x_j) = \mathbb{E}_{x_{\neg j}}\left[F(x_j, x_{\neg j})\right] = \int F(x_j, x_{\neg j})\, dP(x_{\neg j}) \qquad \text{(S6)}$$

where x_j represents the j-th feature of interest, and $x_{\neg j}$ denotes all other features. The function $F(x_j, x_{\neg j})$ refers to the black-box model's prediction based on both x_j and the remaining features. The partial dependence $\widehat{f}_j(x_j)$ is calculated by taking the expected value of the model's output over the distribution of all other features, $x_{\neg j}$, effectively marginalizing their effects.

For a two-way PDP, this process is extended by combining two features, x_j and x_k, where both are simultaneously varied while marginalizing the remaining features. This allows for the visualization of how the combined values of these two features (e.g., temperature and humidity) affect the G/P ratio prediction.

Table S1. Hyperparameters for grid search in random forest model optimization

Hyperparameter	Values
n_estimators	50, 100, 150, 200
max_depth	10, 20, 30, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	sqrt, log2

Table S2. Daytime environmental and gas/particle composition conditions used for predicting G/P ratios of modified monocarboxylic acid in Figure 4

Environmental Conditions	Value
Wind Direction_cos	-0.32
Wind Direction_sin	0.63
Wind Speed (m/s)	1.31
Temperature (°C)	8.74
Humidity (%)	44.81
UV Radiometer A Channel (W m^{-2})	7.18
UV Radiometer B Channel (W m^{-2})	0.12
Photolysis Rate (HONO)(s^{-1})	5.01×10^{-4}
SO_2 (μg m^{-3})	9.55
O_3 (μg m^{-3})	54.48
NO_2 (μg m^{-3})	33.30
NH_3 (μg m^{-3})	2.70
$PM_{2.5}$ (μg m^{-3})	63.07
Cl^- (μg m^{-3})	2.46
SO_4^{2-} (μg m^{-3})	8.09
NO_3^- (μg m^{-3})	25.11
OC (μg m^{-3})	10.04
EC (μg m^{-3})	3.55
NH_4^+ (μg m^{-3})	12.54
K^+ (μg m^{-3})	1.42
pH	4.52
LWC (μg m^{-3})	19.30

Table S3. Nighttime environmental gas/particle composition conditions used for predicting G/P ratios of modified monocarboxylic acid in Figure 4

Environmental Conditions	Value
Wind Direction_cos	-0.19
Wind Direction_sin	0.63
Wind Speed (m/s)	0.93
Temperature (°C)	5.62
Humidity (%)	56.55
UV Radiometer A Channel (W m^{-2})	0
UV Radiometer B Channel (W m^{-2})	0

Photolysis Rate (HONO)(s^{-1})	0
SO_2 (μg m^{-3})	7.55
O_3 (μg m^{-3})	24.80
NO_2 (μg m^{-3})	48.98
NH_3 (μg m^{-3})	2.84
PM2.5 (μg m^{-3})	66.87
Cl^- (μg m^{-3})	2.73
SO_4^{2-} (μg m^{-3})	7.48
NO_3^- (μg m^{-3})	24.84
OC (μg m^{-3})	11.85
EC (μg m^{-3})	4.14
NH_4^+ (μg m^{-3})	12.25
K^+ (μg m^{-3})	1.62
pH	4.40
LWC (μg m^{-3})	28.61

Table S4. Evaluation results of the six single-species models and the selected optimal model parameters

OOMs	R^2	RMSE	n_estimators	max_depth	max_features	min_samples_split
$C_5H_8O_4$	0.88	0.34	200	30	sqrt	2
$C_6H_{10}O_4$	0.85	0.10	50	30	sqrt	2
$C_6H_5NO_3$	0.80	0.03	150	20	sqrt	2
$C_7H_7NO_3$	0.78	0.02	150	30	sqrt	2
$C_{10}H_{16}O_4$	0.54	0.23	200	10	sqrt	2
$C_{12}H_{21}NO_9$	0.51	0.15	100	20	sqrt	10

Table S5. Evaluation results of the six deviation-from-equilibrium G/P ratio models and the selected optimal model parameters

OOMs	R^2	RMSE	n_estimators	max_depth	max_features	min_samples_split
$C_5H_8O_4$	0.83	0.26	50	30	sqrt	2
$C_6H_{10}O_4$	0.79	0.38	100	10	sqrt	2
$C_6H_5NO_3$	0.62	0.40	150	20	sqrt	2
$C_7H_7NO_3$	0.61	0.41	200	10	sqrt	2
$C_{10}H_{16}O_4$	0.53	0.95	100	20	sqrt	2
$C_{12}H_{21}NO_9$	0.52	2.97	200	20	sqrt	5

Table S6. Coefficients in the parameterizations of Mohr et al. (2019) and Ren et al. (2022)

	n_C^0	b_C	b_O	b_{CO}	b_N	Suggested O/C range
Mohr et al. (2019)	25	0.475	0.2	0.9	2.5	-
Ren et al. (2022)	25	0.0700	0.6307	-0.0615	2.3962	0.25–1
Ren et al. (2022)	25	0.2075	2.8276	-1.0744	1.8223	0–0.25

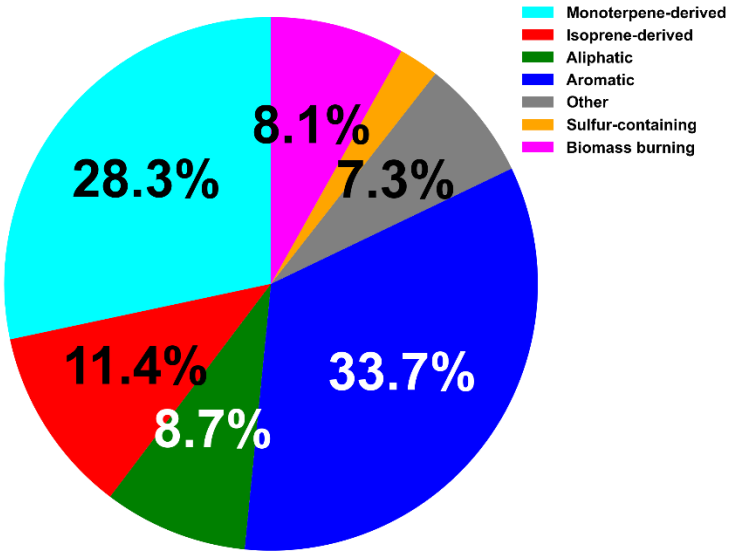

Figure S1. Classification of 123 species used in multi-species model. Sulfur-containing species were directly assigned according to elemental formula. Biomass burning tracer were selected based on literature review. The remaining species were categorized into four precursor types (monoterpene-derived, isoprene-derived, aromatic, and aliphatic) using a random forest model developed by our previous paper (Wang et al., 2024).

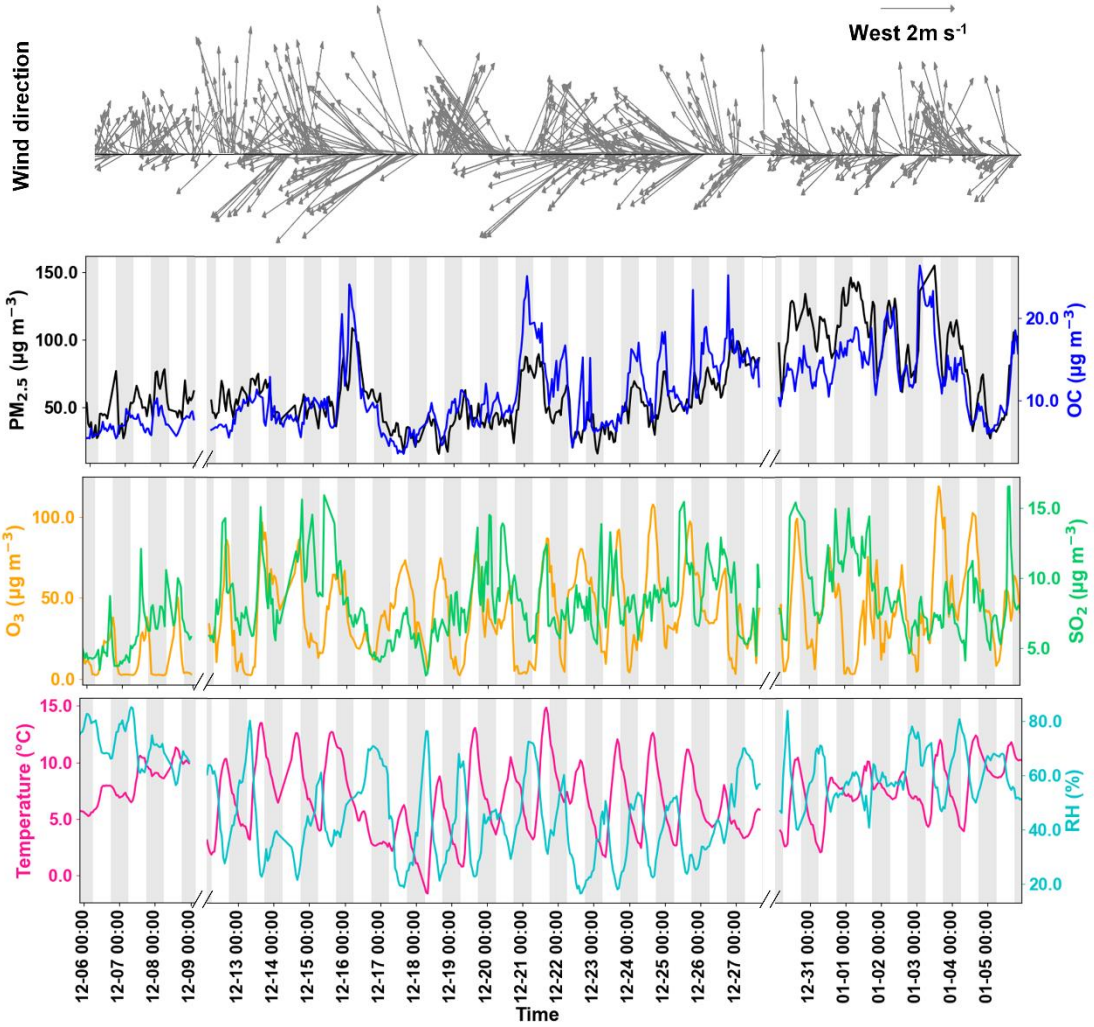

Figure S2. Time series of partial features. (a) Wind direction and wind speed (reference vector: west 2 m s^{-1}). (b) $PM_{2.5}$ and organic carbon (OC) concentrations. (c) Ozone (O_3) and sulfur dioxide (SO_2) concentrations. (d) Ambient temperature and relative humidity (RH).

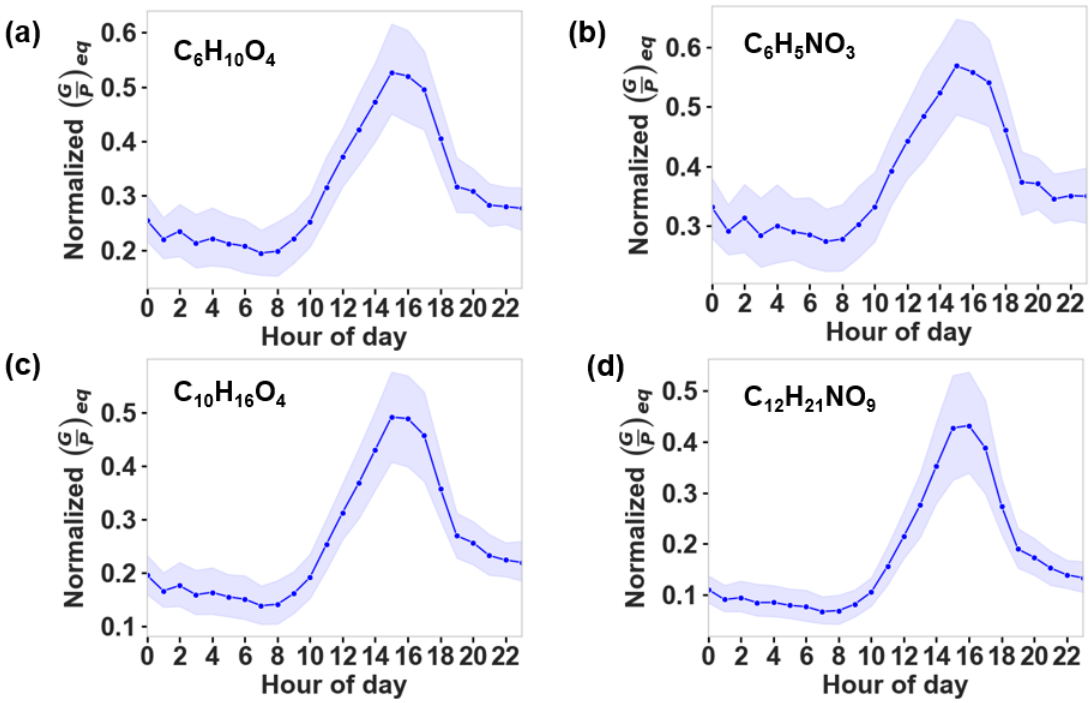

Figure S3. Diurnal variations of normalized equilibrium G/P ratios for(a) Adipic acid $C_6H_{10}O_4$, (b) Nitrophenol $C_6H_5NO_3$, (c) Monoterpene oxidation product $C_{10}H_{16}O_4$ and (d) Nitrated aliphatic acid $C_{12}H_{21}NO_9$. Blue shade denotes 95% confidence interval.

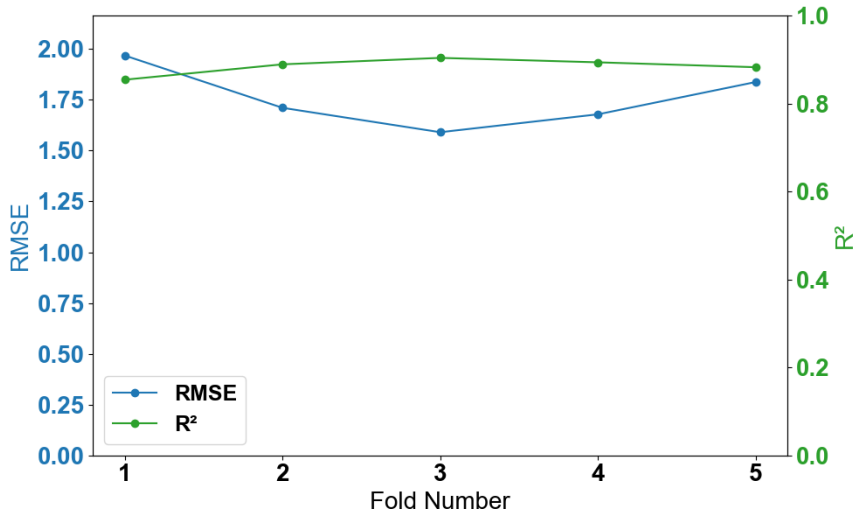

Figure S4. RMSE and R^2 values across five-fold cross-validation for the multi-species model, with RMSE (left axis) and R^2 (right axis) shown for each fold.

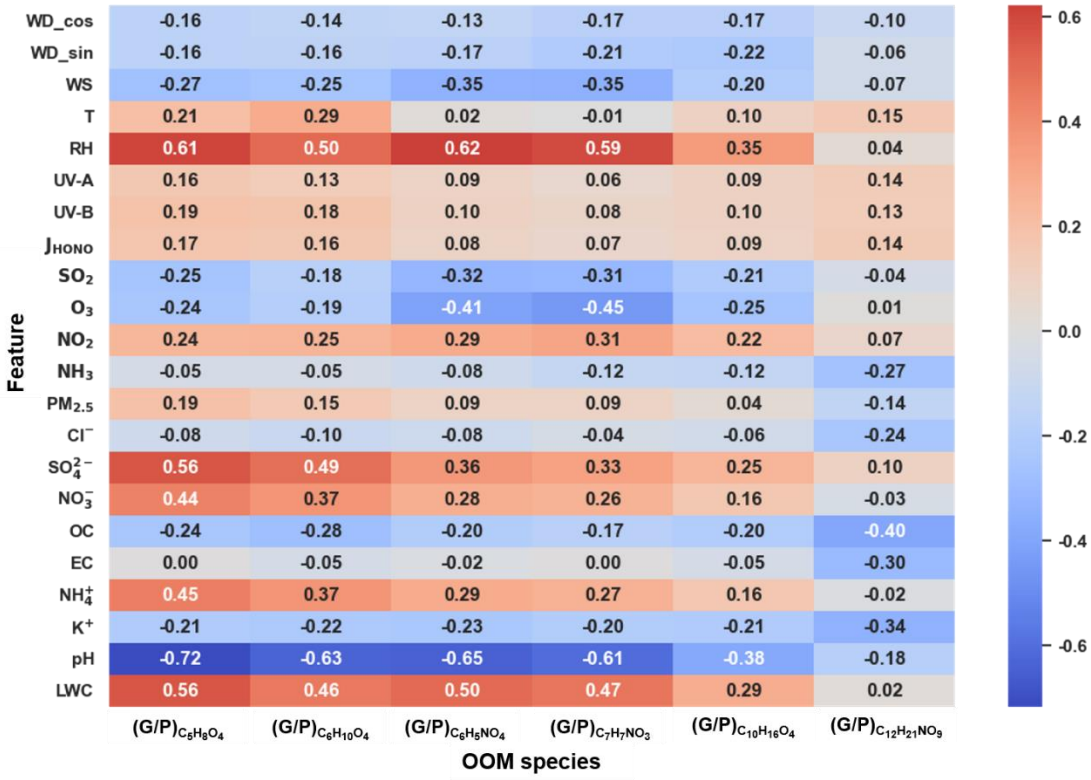

Figure S5. The correlation between the observed G/P ratios of the 6 selected species and various features.

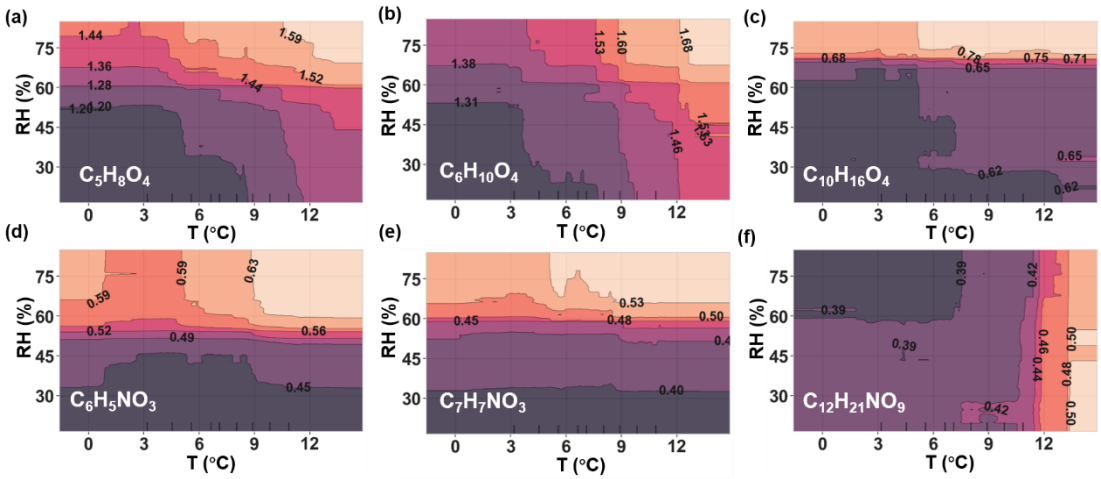

Figure S6. Two-way partial dependence plots showing the impact of temperature and humidity on the predicted G/P ratio for (a) Glutaric acid ($C_5H_8O_4$), (b) Adipic acid ($C_6H_{10}O_4$), (c) Monoterpene oxidation product ($C_{10}H_{16}O_4$), (d) Nitrophenol ($C_6H_5NO_3$), (e) Nitrophenol ($C_7H_7NO_3$), and (f) Nitrated aliphatic acid ($C_{12}H_{21}NO_9$). The numbers on the contour lines denote the predicted G/P ratios.

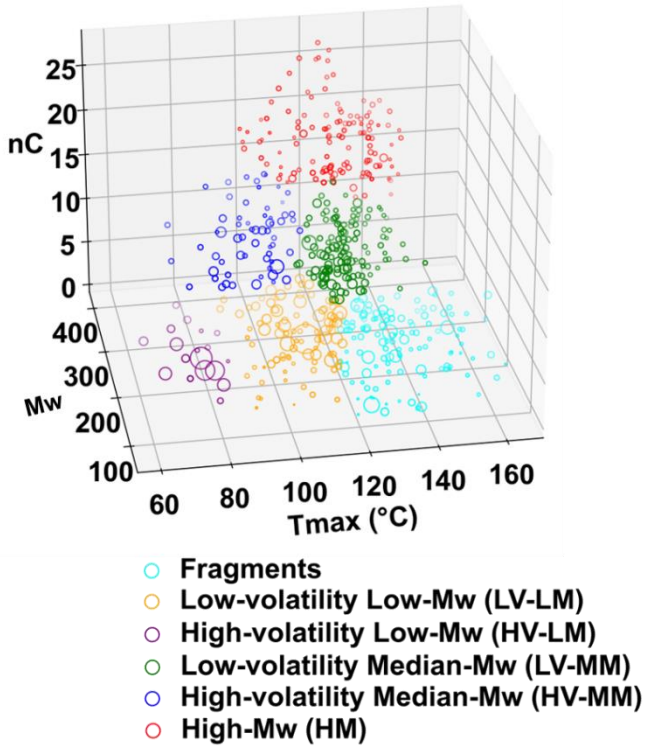

Figure S7. Groups of OOMs identified using K-means clustering analysis, showing the three-dimensional distribution of these species by T_{max}, molecular weight, carbon number, and concentration. The size of each circle is proportional to the concentration of the corresponding OOMs.

References

Faxon, C., Hammes, J., Le Breton, M., Pathak, R. K., and Hallquist, M.: Characterization of organic nitrate constituents of secondary organic aerosol (SOA) from nitrate-radical-initiated oxidation of limonene using high-resolution chemical ionization mass spectrometry, Atmos. Chem. Phys., 18, 5467-5481, https://doi.org/10.5194/acp-18-5467-2018, 2018.

Fountoukis, C. and Nenes, A.: ISORROPIA II: a computationally efficient thermodynamic equilibrium model for K^+–Ca^{2+}–Mg^{2+}–NH_4^+–Na^+–SO_4^{2-}–NO_3^-–Cl^-–H_2O aerosols, Atmos. Chem. Phys., 7, 4639-4659, https://doi.org/10.5194/acp-7-4639-2007, 2007.

Guo, H., Sullivan, A. P., Campuzano-Jost, P., Schroder, J. C., Lopez-Hilfiker, F. D., Dibb, J. E., Jimenez, J. L., Thornton, J. A., Brown, S. S., Nenes, A., and Weber, R. J.: Fine particle pH and the partitioning of nitric acid during winter in the northeastern United States, J. Geophys. Res. Atmos., 121, 10,355-310,376, https://doi.org/10.1002/2016JD025311, 2016.

Jia, Y., Hu, X., Kang, W., and Dong, X.: Unveiling Microbial Nitrogen Metabolism in Rivers using a Machine Learning Approach, Environ. Sci. Technol., 58, 6605-6615, https://doi.org/10.1021/acs.est.3c09653, 2024.

Lundberg, S.: A unified approach to interpreting model predictions, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, Dec 04-09, 2017, https://doi.org/10.48550/arXiv.1705.07874, 2017.

Mohr, C., Thornton, J. A., Heitto, A., Lopez-Hilfiker, F. D., Lutz, A., Riipinen, I., Hong, J., Donahue, N. M., Hallquist, M., Petäjä, T., Kulmala, M., and Yli-Juuti, T.: Molecular identification of organic vapors driving atmospheric nanoparticle growth, Nat. Commun., 10, 4442, https://doi.org/10.1038/s41467-019-12473-2, 2019.

Peräkylä, O., Riva, M., Heikkinen, L., Quéléver, L., Roldin, P., and Ehn, M.: Experimental investigation into the volatilities of highly oxygenated organic molecules (HOMs), Atmos. Chem. Phys., 20, 649-669, https://doi.org/10.5194/acp-20-649-2020, 2020.

Priestley, M., Kong, X., Pei, X., Pathak, R. K., Davidsson, K., Pettersson, J. B. C., and Hallquist, M.: Volatility Measurements of Oxygenated Volatile Organics from Fresh and Aged Residential Wood Burning Emissions, ACS Earth Space Chem., 8, 159-173, https://doi.org/10.1021/acsearthspacechem.3c00066, 2024.

Ren, S., Yao, L., Wang, Y., Yang, G., Liu, Y., Li, Y., Lu, Y., Wang, L., and Wang, L.: Volatility parameterization of ambient organic aerosols at a rural site of the North China Plain, Atmos. Chem. Phys., 22, 9283-9297, https://doi.org/10.5194/acp-22-9283-2022, 2022.

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, California, USA, Aug 13-17, 2016, https://doi.org/10.1145/2939672.2939778, 2016.

Shi, H., Yang, N., Yang, X., and Tang, H.: Clarifying Relationship between PM2.5 Concentrations and Spatiotemporal Predictors Using Multi-Way Partial Dependence Plots, Remote Sens., 15, 358, https://doi.org/10.3390/rs15020358, 2023.

Stark, H., Yatavelli, R. L. N., Thompson, S. L., Kang, H., Krechmer, J. E., Kimmel, J. R., Palm, B. B., Hu, W. W., Hayes, P. L., Day, D. A., Campuzano-Jost, P., Canagaratna, M. R., Jayne, J. T., Worsnop, D. R., and Jimenez, J. L.: Impact of Thermal Decomposition on Thermal Desorption Instruments: Advantage of Thermogram Analysis for Quantifying Volatility Distributions of Organic Species, Environ. Sci. Technol., 51, 8491-8500, https://doi.org/10.1021/acs.est.7b00160, 2017.

Wang, X., Zhao, Y., Hu, K., Wang, J., Wang, Q., Chen, N., Zhu, B., Zhang, H.-H., and Yu, H.: Linking Precursors and Volatility of Ambient Oxygenated Organic Aerosols Using Thermal Desorption Measurement and Machine Learning, ACS ES&T Air, 1, 1239-1251, https://doi.org/10.1021/acsestair.4c00076, 2024.