A point-to-point response and relevant changes made in the revised manuscript

Ms. Ref. No.: egusphere-2025-229

Title: Investigating Influencing Factors of Gas-Particle Distribution of Oxygenated Organic Molecules in Urban Atmosphere and Its Deviation from Equilibrium Partitioning Using Random Forest Model.

Anonymous Referee #1

Wang et al. measured both gaseous and particle-phase OOMs using FIGAERO-CIMS during a winter campaign in Wuhan. They derived gas-to-particle ratios (G/P) using measured FIGAERO-CIMS signals and predicted equilibrium G/P based on predicated OOM volatility. They further applied machine learning methods to revealed key factors that are associated with G/P. The manuscript aligns with the scope of Atmospheric Chemistry and Physics. However, clarification of the principal findings is recommended prior to publication. Specific comments are given below:

Response:

We sincerely thank the reviewer for their valuable comments and suggestions, which will significantly improve the quality of our manuscript.

1. *The machine learning analysis provides intriguing insights into G/P influencing factors. However, given potential limitations in the representativeness of the dataset and the ML methodology, the mechanistic interpretation of identified factors may not be entirely clear. I recommend expanded discussion in Section 3.2.2 to include detailed analysis of at least one parameter (e.g., temperature).*

Response:

By itself, the machine learning methodology cannot uncover the fundamental mechanisms through which the factors influence G/P ratio. In Section 3.2.2, we (1) examined the positive or negative effects and the sensitive ranges of key features like RH, LWC, OC, K^+, SO_4^{2-} and pH identified by the ML model and (2) explained why ambient temperature did not rank as important feature for most of the OOM species. Prior publications were provided to support our interpretation.

We expanded our discussion of the impact of temperature, RH and LWC. In lines 398-402,

"One explanation is that the low RH and LWC in particles may facilitate the reversible formation of oligomers (Shen et al. (2018) and suppress their hydrolysis (Liu et al., 2012), thereby increasing the concentration of these OOMs in particle phase. It is also possible that the thermal desorption and subsequent detection of particle-bound OOMs were inhibited in aerosol particles with more moisture."

In lines 428-440,

"In general, temperature is supposed to be an important influential factor of G/P ratio, because saturation vapor pressure of OOMs increases with temperature. Temperature ranged from -

1.6 °C to 14.9 °C during the winter campaign. Although temperature increase tends to elevate the G/P ratios as expected (Figure 6a), the models show that temperature did not rank as important feature for 5 out of the 6 OOM species. We evaluated the effect of temperature on G/P ratios using two-way partial dependence plots (Figure S6). G/P ratio is sensitive to temperature change only for two dicarboxylic acids ($C_5H_8O_4$ and $C_6H_{10}O_4$, Figure S6a-S6b) and for $C_{12}H_{21}NO_9$ in a narrow temperature range of 10-13 °C (Figure S6f and Figure 7h). The G/P ratios of $C_6H_5NO_3$, $C_7H_7NO_3$ and $C_{10}H_{16}O_4$ are not sensitive to temperature across most of the RH range. This behavior may be attributed to the aerosol coating of inorganic salts and other aerosol components that hinder the rapid equilibrium partitioning of OOMs when temperature changes. In addition, the influence of temperature may be obscured due to the dominant effect of particle composition features (e.g., LWC, pH, OC, SO_4^{2-}, and K^+) as discussed above."

2. *Please explain Eq. 4 with an emphasis on its underlying assumptions that are possibly violated in the real atmosphere. This clarification would aid the discussion on the influencing factors of the ratio of $(G/P)_{obs}$ to $(G/P)_{eq}$.*

Response:

Equation 4 is primarily based on Raoult's Law. In line 143-150, we explain the derivation of Eq 4 and underlying assumptions.

"According to modified Raoult's Law, the saturation ratio of an organic species in gas phase (i.e. $\frac{C_g}{C^*}$) equals the mass fraction of the species in organic aerosol with mass concentration C_{OA} ($i.e.\frac{C_p}{C_{OA}}$), under the assumptions of equilibrium absorptive partitioning of the species over an ideal organic solution and that the species has a molecular weight similar to that of the organic solution (Donahue et al., 2009; Epstein et al., 2010). The equilibrium G/P ratio $(\frac{G}{P})_{eq}$ can thus be estimated from saturated mass concentration C^* and mass concentration of organic aerosol C_{OA} ($C_{OA} = C_{OC} \times 1.4$) using Eq. (4)

$$(\frac{G}{P})_{eq} = \frac{C^*(T)}{C_{OA}} \qquad (4)$$ "

3. *Lines 179-181, Page 7. Please specify the data partitioning strategy for training and test sets and the measures to prevent model overfitting.*

Response:

Thank you for the reviewer's suggestion. Regarding the data partitioning strategy for the training and test sets, we performed a random split. In lines 207-208, we revised the text to:

"The data used for modeling were randomly divided into training data (85% of the total) for model training and test data (15% of the total) for evaluating model generalization."

Unlike gradient boosting regression and neural networks, the random forest algorithm trains multiple decision trees using bootstrapped subsets of the training data and random subsets of features, which inherently provides better resistance to overfitting (Amaratunga et al., 2008).

To further prevent model overfitting and enhance its generalization ability, we implemented the following 3 measures:

1) Cross-validation: We employed 5-fold cross-validation on the training set to assess the model's generalization ability on different subsets of data. We revised the text in lines 242-243 to:

"After selecting the optimal hyperparameters, we further evaluated the final model using 5-fold cross-validation to assess the model's generalization ability and ensure it was not overfitted."

2) Restricting parameter ranges: In the Grid Search method, we restricted the parameter ranges for n_estimators, max_depth, max_features, and min_samples_split to prevent overfitting. In lines 239-241, we added the following sentence:

"The specific hyperparameter settings used in the Grid Search are provided in the supplementary materials, in Table S1."

In the supplementary materials, we have added Table S1:

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

3) Evaluation of Model Generalization Using the Test Set: The test set, which was not involved in the training process, provides a reliable assessment of the model's generalization ability, helping to prevent overfitting. The R^2 evaluations presented in Table S4 and Table S5 of the supplementary materials are based on the test set results, demonstrating the model's satisfactory generalization performance.

In lines 330-332, we revised the sentence to:

"The 5-fold cross-validation demonstrates that a predictive multi-species model with satisfactory generalization performance was developed, achieving R^2=0.88 ± 0.02 and RMSE = 1.76 ± 0.13 on the test set (Figure S4)."

In lines 374-376, we revised the sentence to:

"The evaluation results and optimal parameters of the six single-species models on the test set are presented in Table S4. All models show acceptable generalization ability (R^2 = 0.51-0.88)."

In lines 458-459, we revised the sentence to:

"All the models show acceptable generalization performance (R^2 = 0.52-0.83) (Table S5) on the test set."

4. *Fig. 1 and its relevant discussion. The method in Ren et al. (2022) provided the equilibrium*

pressure (C^{eq}) rather than the saturation pressure (C^*), as the formula was fit to atmospheric aerosols (a mixture of many OOMs). Using C^{eq} instead of C^* in Eq. 4 may introduce systematic biases. Could this be the reason for the observed discrepancy between $(G/P)_{obs}$ and $(G/P)_{eq}$ in Figs 1a and 1d?

Response:

In Ren et al. (2022), the authors first obtained a calibration curve for C^* and T_{max} using a series of polyethylene glycol standards with known saturation vapor pressures. Then they measured T_{max} values of a number of OOMs in atmospheric aerosols and converted the T_{max} values into corresponding C^* of the OOMs. Therefore, Ren et al. indeed provided the saturation concentration (C^*) values of OOMs with known formulas, although the OOMs were obtained from ambient aerosols in their experiment.

5. *Previous studies (e.g., Voliotis et al., 2021; Chen et al., 2024) have reported a narrow volatility range of OOMs retrieved using the partitioning method. It is not surprising to see a huge difference between the volatility obtained using different methods. Would it be possible to expand discussion on this finding?*

Response:

Thank you for providing the relevant literature. We cited them and expanded the discussion in lines 295-306 as follows:

"Among all the predictions, the prediction from Priestley et al. (2024) is most close to our observation. This is because their C^* parameterization is based on the measured gas and particle-phase concentrations of OOMs in fresh or aged residential wood-burning emissions. Their predicted G/P ratio is thus inherently consistent with the observed G/P ratios in our study.

This also highlights the risks of estimating volatility (C^*) using the partitioning method, which is based on measuring equilibrium gas- and particle-phase concentrations of OOMs. Two key issues arise: (1) OOMs may not achieve the assumed equilibrium state in real atmospheric or chamber conditions, introducing substantial uncertainty into calculations of $(\frac{G}{P})_{eq}$; (2) The method fails for the compounds with extremely high or low volatility, as their gas- or particle-phase concentrations often fall below the detection limit of mass spectrometers. These limitations explain why the partitioning method typically reports a narrow volatility range (Voliotis et al., 2021; Chen et al., 2024)."

6. *As noted by the authors, atmospheric OOMs may not reach equilibrium between the gas and particle phases (e.g., Li et al., 2024). Could machine learning features capture this non-equilibrium effects?*

Response:

Yes, machine learning features could capture the non-equilibrium effects. This was discussed in Section 3.3 "Identifying key factors driving the deviations of gas/particle partitioning from equilibrium state". The ML model identified RH, LWC, O_3 and temperature as four features that lead to non-equilibrium partitioning.

In lines 44-45, we add the citation Li et al., 2024:

"As a result, OOMs rarely achieve equilibrium partitioning between the gas and particle phases (Roldin et al., 2014; Li et al., 2024)."

References

Amaratunga, D., Cabrera, J., and Lee, Y.-S.: Enriched random forests, Bioinformatics, 24, 2010-2014, http://doi.org/10.1093/bioinformatics/btn356, 2008.

Chen, W., Hu, W., Tao, Z., Cai, Y., Cai, M., Zhu, M., Ye, Y., Zhou, H., Jiang, H., Li, J., Song, W., Zhou, J., Huang, S., Yuan, B., Shao, M., Feng, Q., Li, Y., Isaacman-VanWertz, G., Stark, H., Day, D. A., Campuzano-Jost, P., Jimenez, J. L., and Wang, X.: Quantitative Characterization of the Volatility Distribution of Organic Aerosols in a Polluted Urban Area: Intercomparison Between Thermodenuder and Molecular Measurements, J. Geophys. Res. Atmos., 129, e2023JD040284, https://doi.org/10.1029/2023JD040284, 2024.

Donahue, N. M., Robinson, A. L., and Pandis, S. N.: Atmospheric organic particulate matter: From smoke to secondary organic aerosol, Atmospheric Environment, 43, 94-106, https://doi.org/10.1016/j.atmosenv.2008.09.055, 2009.

Epstein, S. A., Riipinen, I., and Donahue, N. M.: A Semiempirical Correlation between Enthalpy of Vaporization and Saturation Concentration for Organic Aerosol, Environ. Sci. Technol., 44, 743-748, https://doi.org/10.1021/es902497z, 2010.

Li, Y., Cai, R., Yin, R., Li, X., Yuan, Y., An, Z., Guo, J., Stolzenburg, D., Kulmala, M., and Jiang, J.: A kinetic partitioning method for simulating the condensation mass flux of organic vapors in a wide volatility range, J. Aerosol Sci., 180, 106400, https://doi.org/10.1016/j.jaerosci.2024.106400, 2024.

Liu, S., E., S. J., Chen, S., Naruki, H., A., Z. R., and and Russell, L. M.: Hydrolysis of Organonitrate Functional Groups in Aerosol Particles, Aerosol Sci. Technol., 46, 1359-1369, http://doi.org/10.1080/02786826.2012.716175, 2012.

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.

Roldin, P., Eriksson, A. C., Nordin, E. Z., Hermansson, E., Mogensen, D., Rusanen, A., Boy, M., Swietlicki, E., Svenningsson, B., Zelenyuk, A., and Pagels, J.: Modelling non-equilibrium secondary organic aerosol formation and evaporation with the aerosol dynamics, gas- and particle-phase chemistry kinetic multilayer model ADCHAM, Atmos. Chem. Phys., 14, 7953-7993, https://doi.org/10.5194/acp-14-7953-2014, 2014.

Shen, H., Chen, Z., Li, H., Qian, X., Qin, X., and Shi, W.: Gas-Particle Partitioning of Carbonyl Compounds in the Ambient Atmosphere, Environ. Sci. Technol., 52, 10997-11006, http://doi.org/10.1021/acs.est.8b01882, 2018.

Voliotis, A., Wang, Y., Shao, Y., Du, M., Bannan, T. J., Percival, C. J., Pandis, S. N., Alfarra, M. R., and

McFiggans, G.: Exploring the composition and volatility of secondary organic aerosols in mixed anthropogenic and biogenic precursor systems, Atmos. Chem. Phys., 21, 14251-14273, https://doi.org/10.5194/acp-21-14251-2021, 2021.