Supplementary Information

Differentiation of primary and secondary marine organic aerosol with machine learning

Baihua Chen[1], Lu Lei[2], Emmanuel Chevassus[2], Wei Xu[1*], Ling Zhen[1], Haobin Zhong[1], Lin Wang[1], Chunshui Lin[3], Ru-Jin Huang[3], Darius Ceburnis[2], Colin O'Dowd[2], Jurgita Ovadnevaite[2*]

[1]State Key Laboratory of Advanced Environmental Technology, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China.
[2]School of Physics, Centre for Climate & Air Pollution Studies, Ryan Institute, University of Galway, Galway, Ireland.
[3]State Key Laboratory of Loess and Quaternary Geology and Key Laboratory of Aerosol Chemistry and Physics, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China.

Correspondence to: Wei Xu (wxu@iue.ac.cn) and Jurgita Ovadnevaite (jurgita.ovadnevaite@universityofgalway.ie)

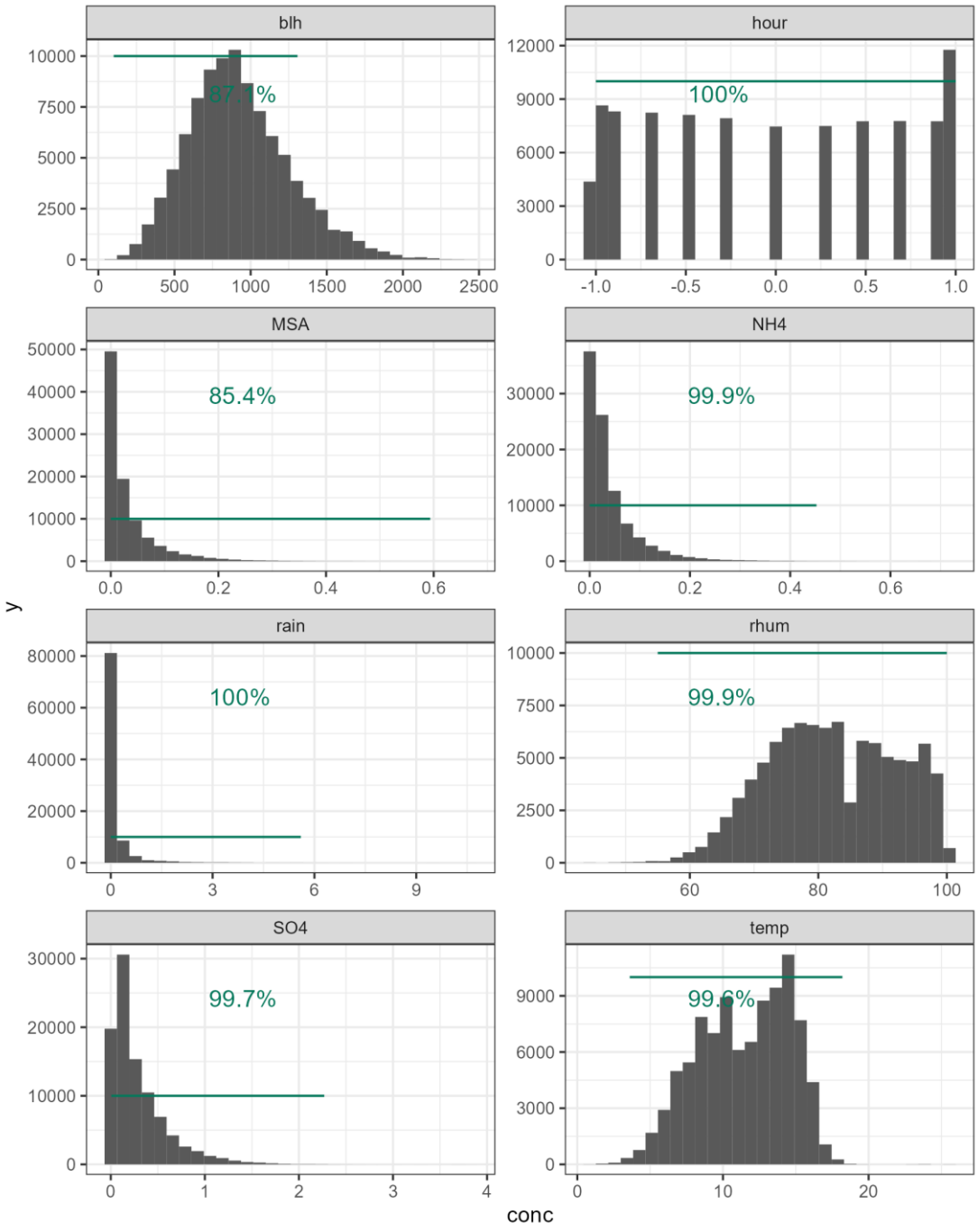


Figure S1. Frequency distribution of each predictors across the entire clean marine dataset and the range of the data used in SOA model training, the label indicates the data coverage.

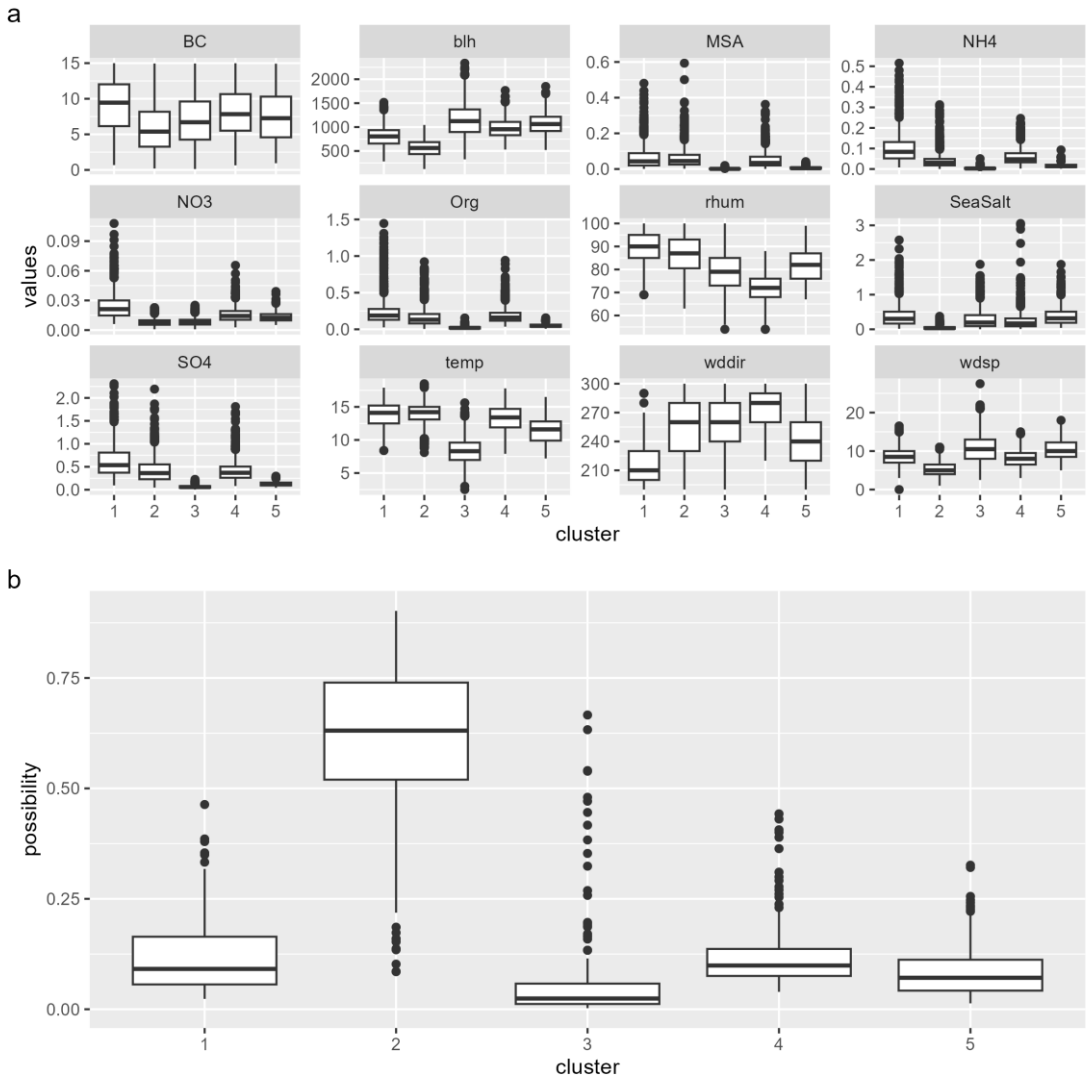


Figure S2. (a) Factor Profile Resolution via Fuzzy C-Means Clustering: This figure displays the profiles of factors resolved using fuzzy c-means clustering. To enhance clarity and relevance, only data with a cluster membership confidence exceeding 80% are included. This visualization helps in identifying the predominant characteristics associated with each factor. (b) Probability of the selected training data attribution to each cluster: This panel illustrates the likelihood that selected training data correspond to each factor identified by the clustering analysis. Notably, the second factor emerges as the most probable secondary factor, characterized by high concentrations of non-sea salt sulfate (nss-SO_4) and methanesulfonic acid (MSA), alongside low levels of sea salt and wind speed. The data indicate that training selections are most frequently associated with this 2nd factor, suggesting its significant role in the composition of the dataset and its potential impact on the model's training efficacy.

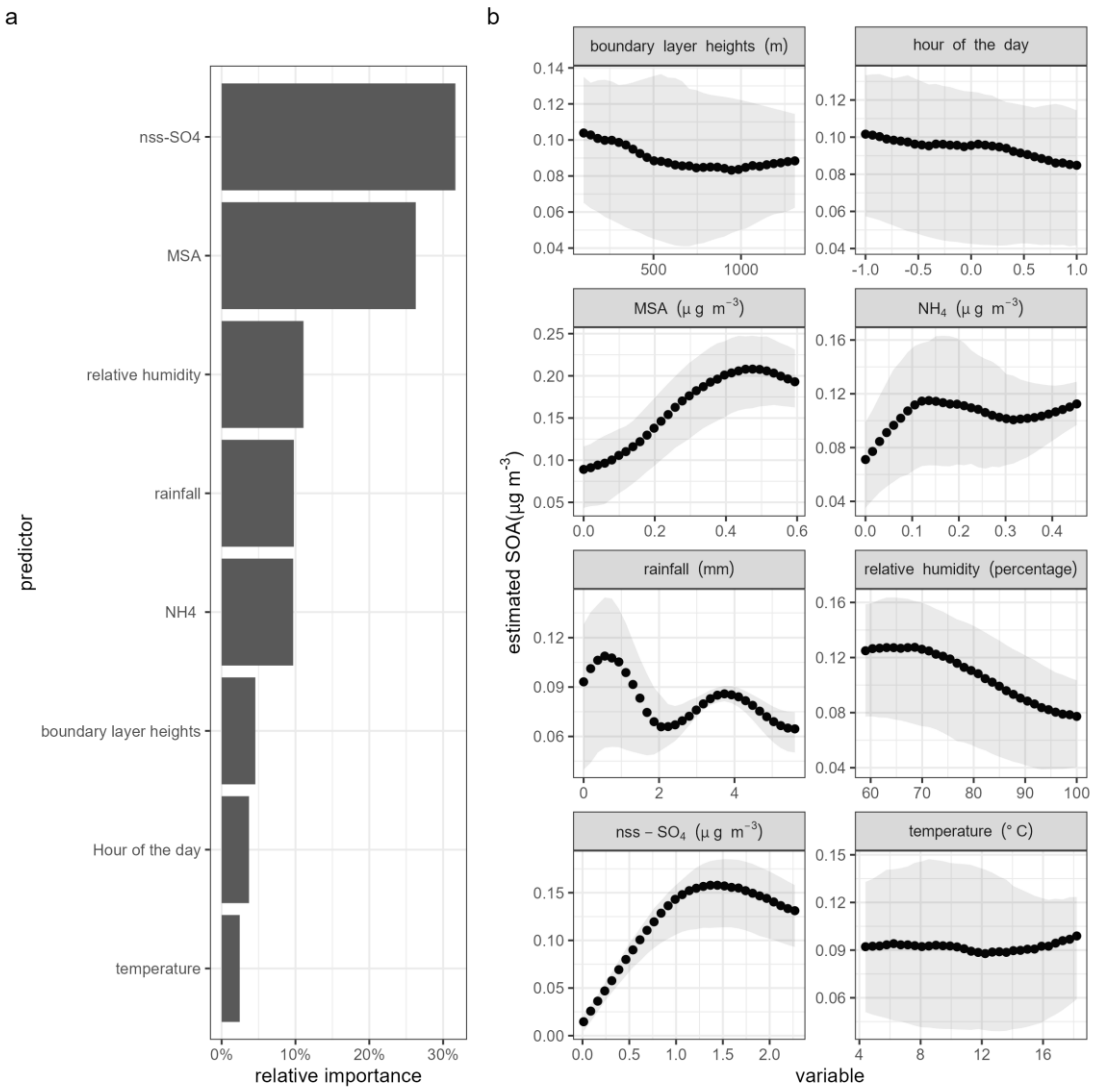


Figure S3. (a) the relative importance of different predictors. The hour and month were transformed with cosine functions. The month and hour were transformed to preserve continuity in the data. (b) The partial dependence of estimated SOA to different predictors. The lines represent median values and the shaded area represent 25th to 75th percentile; For example, the 23:00pm and 01:00 am are numerically far apart while they are temporally close. Therefore we transformed the hour of the day to periodic functions simulated with cosine functions as follows: hour = cos(hour*(2pi/12).

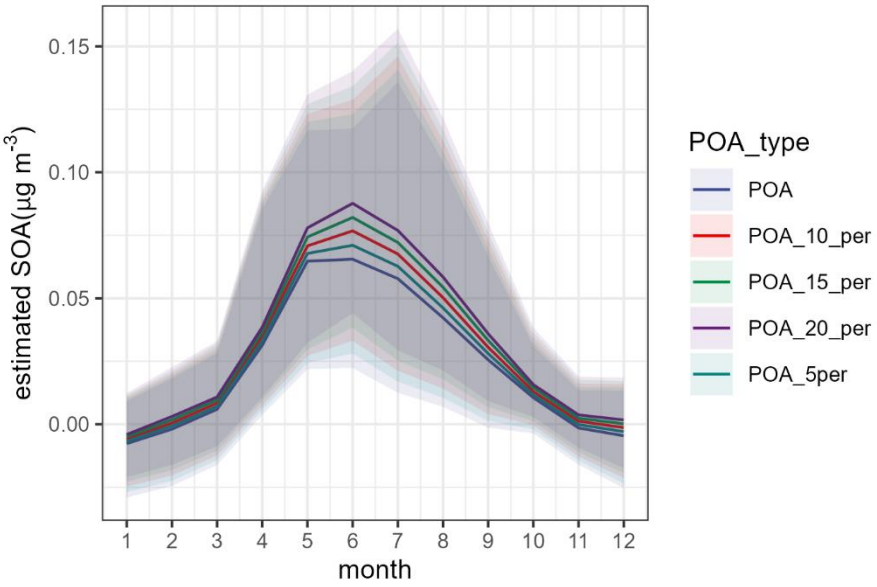


Figure S4. POA seasonality with different sensitivity test. Assuming the POA contribution (5%, 10%, 15% and 20%) to the total OA in selected secondary marine aerosol data. The shaded area represents 25th to 75th percentile of each test. Assuming that POA accounts for different percentages of total OA is likely to produce many negative predictions, especially in wintertime.

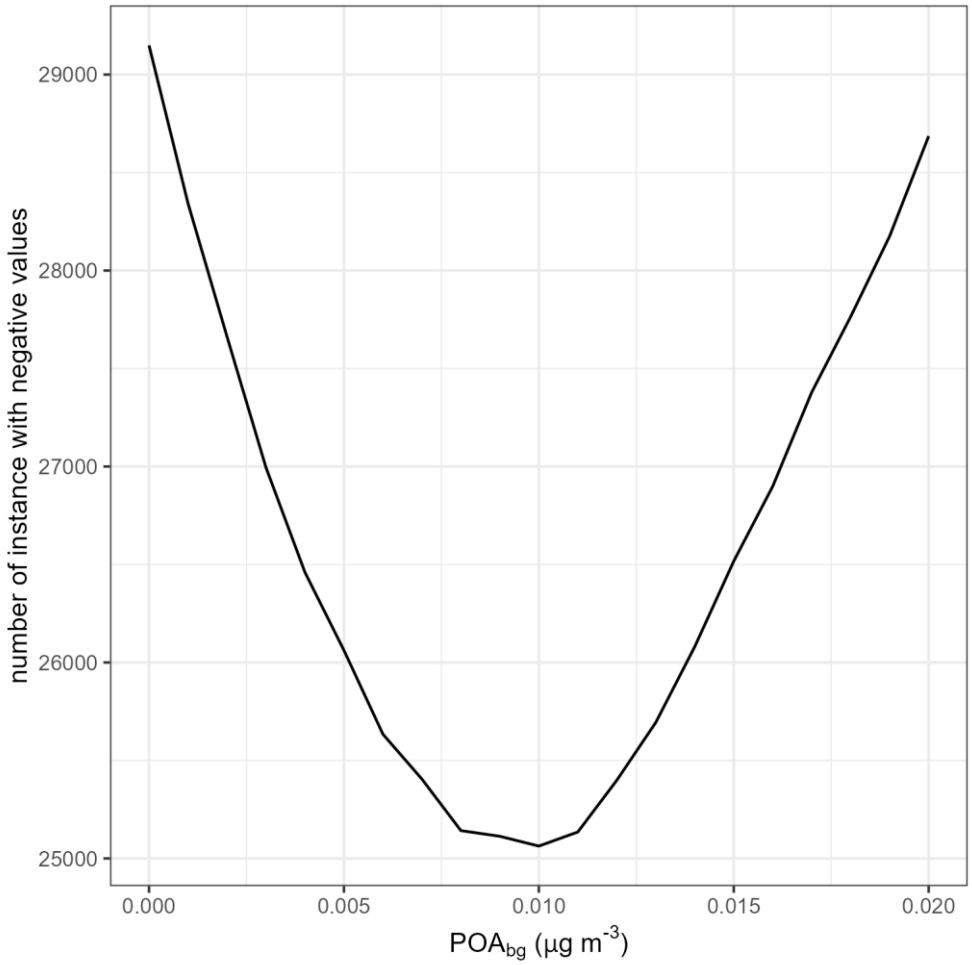


Figure S5. The different POA_{bg} value used in ML model and the number of non-physical prediction instance including POA and SOA (lower than 0). This indicates the POA_{bg} of 0.01 µg m^{-3} is the optimum value.

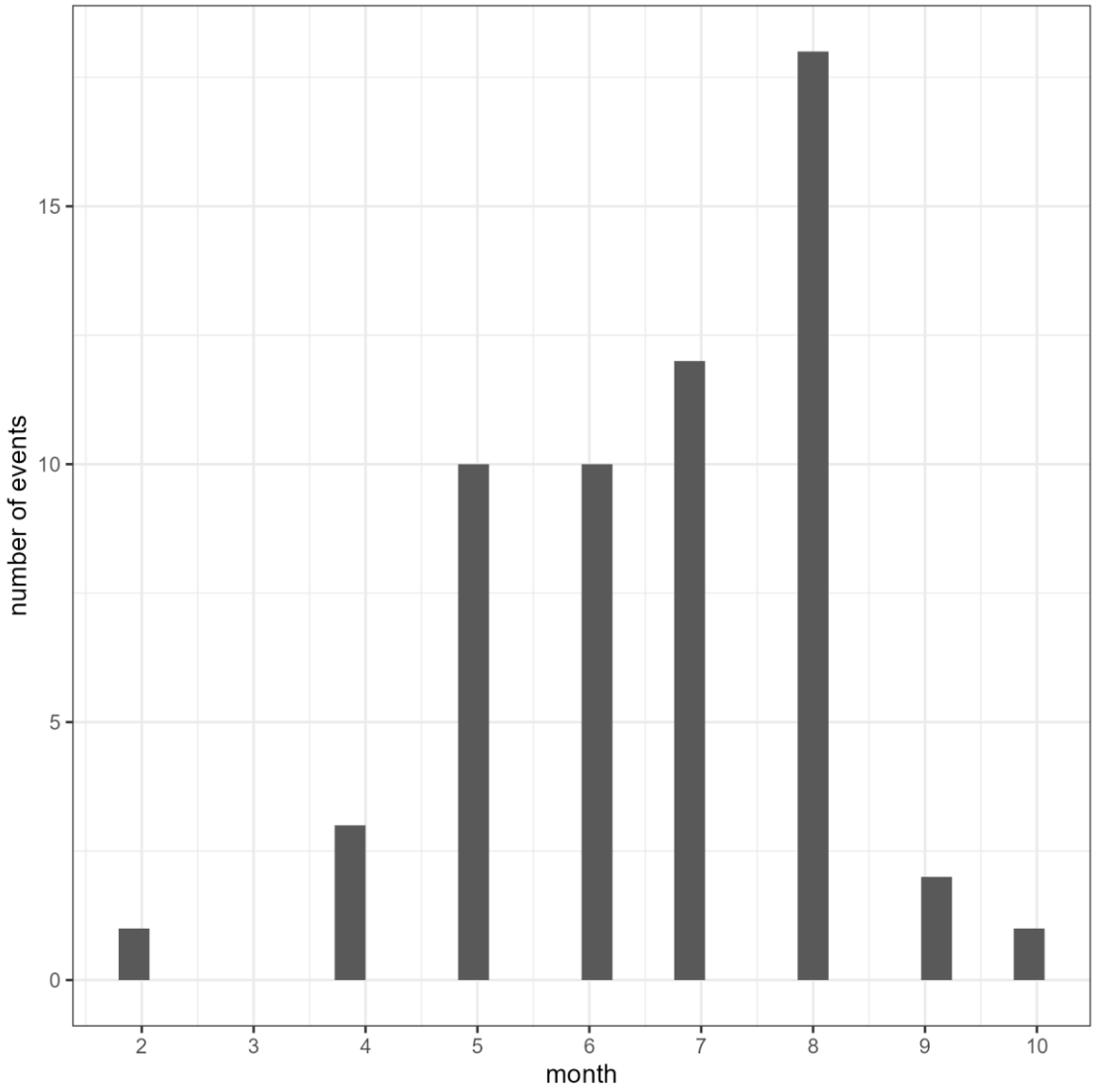


Figure S6. The monthly distribution of POA events. POA event is defined as a period with POA concentration higher than 0.1 μg m^{-3} over 12 hours.

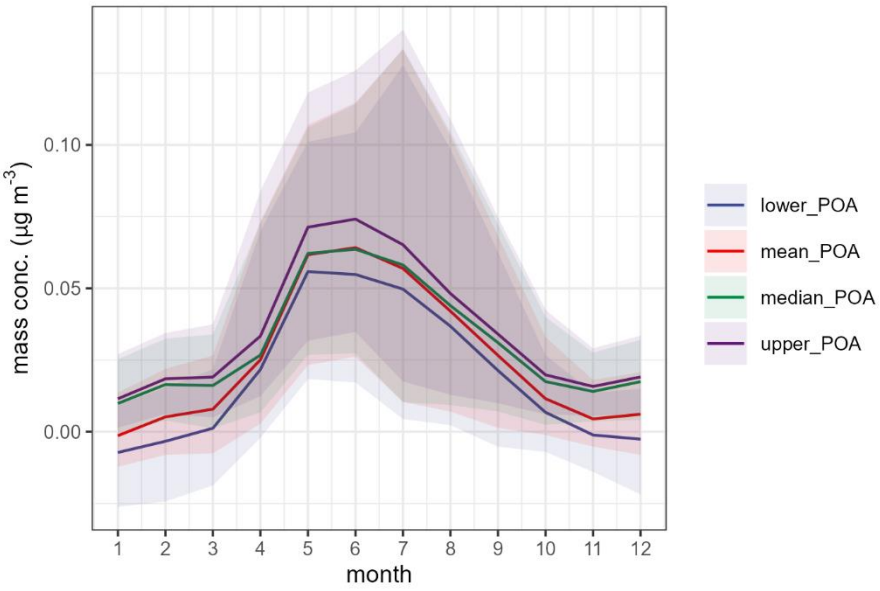


Figure S7. Monte Carlo simulations. The Monte Carlo simulation is done by randomly drop 20% of the data in training dataset and repeat the training process for 1000 times. The mean, median, lower (25th percentile) and upper (75th percentile) of the simulation ensemble were summarised monthly.

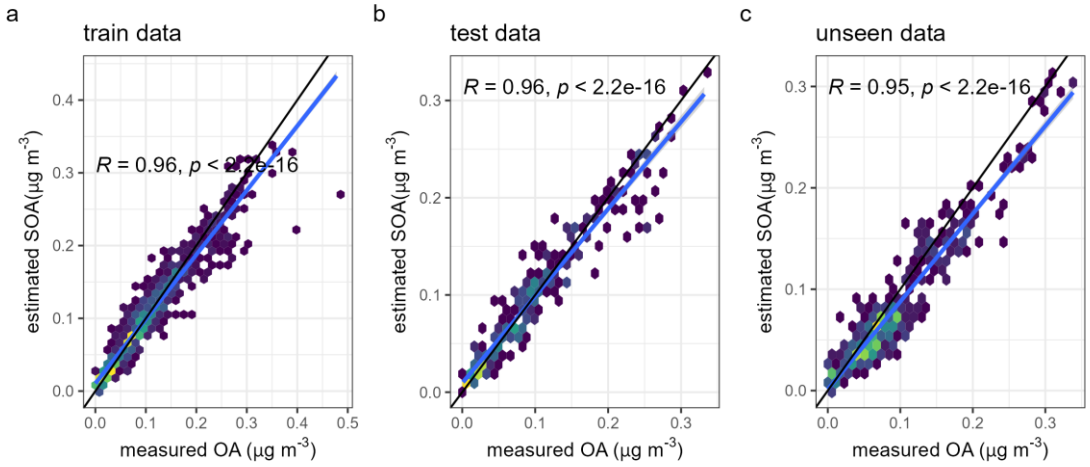


Figure S8. The performance of the ML models without MSA as predictors.

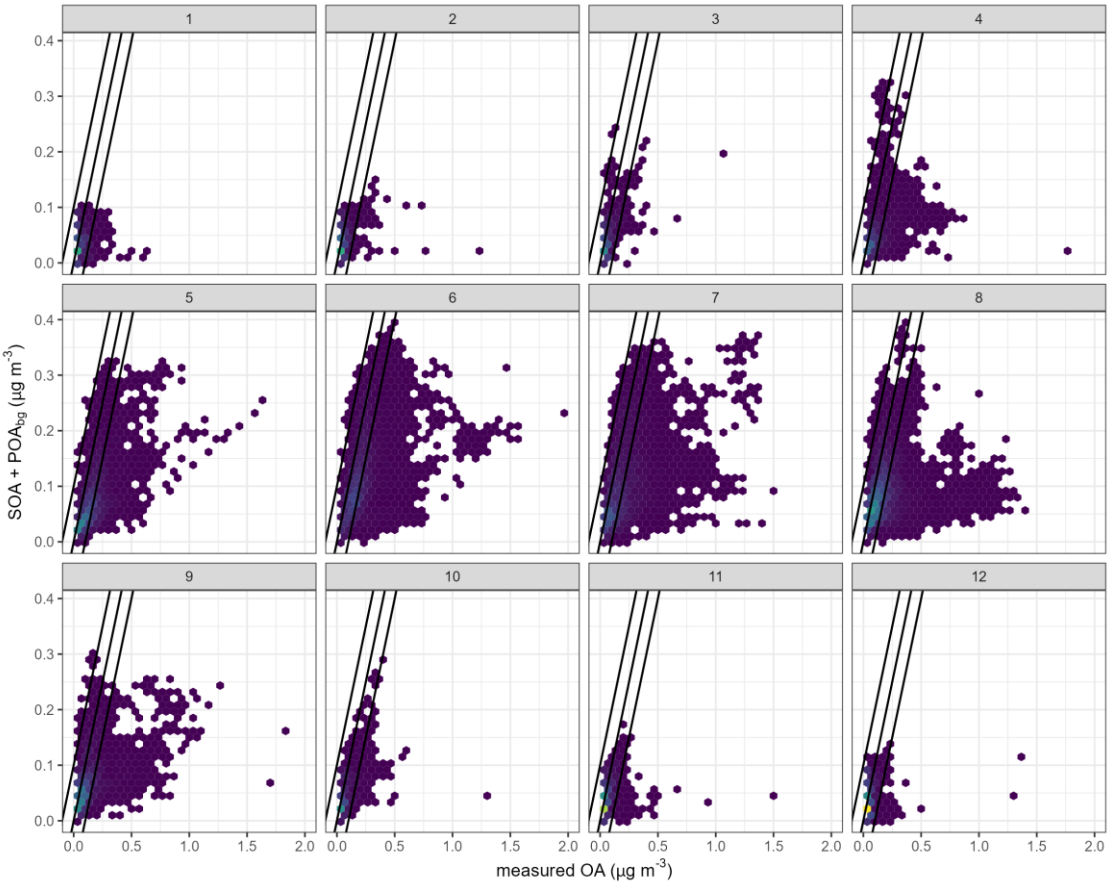
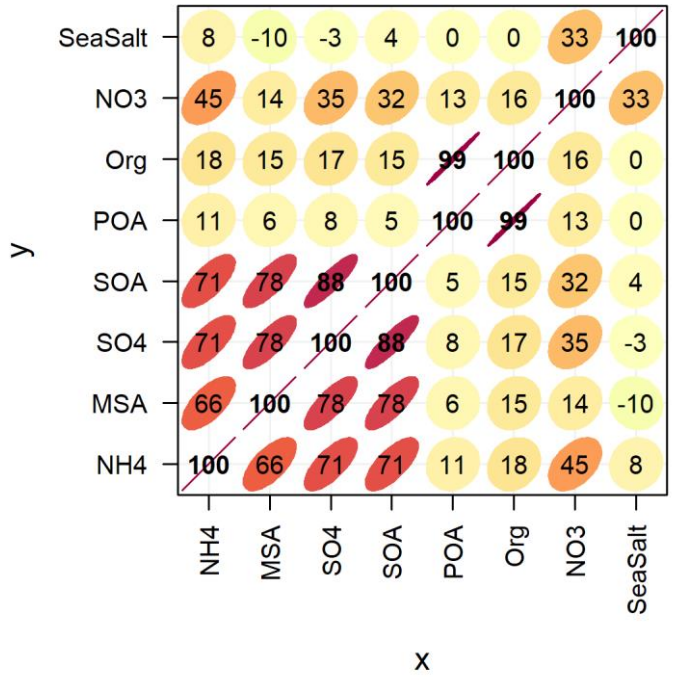


Figure S9. The estimated SOA + POA_{bg} versus measured total OA in different months. The black lines represent y = x - 0.1, y = x, and y = x+0.1.



60 **Figure S10. Correlation between chemical species for the entire clean marine air mass data.**