Supplement of Atmos. Chem. Phys., 25, 14205–14219, 2025
https://doi.org/10.5194/acp-25-14205-2025-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Differentiation of primary and secondary marine organic aerosol with machine learning

Baihua Chen et al.

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

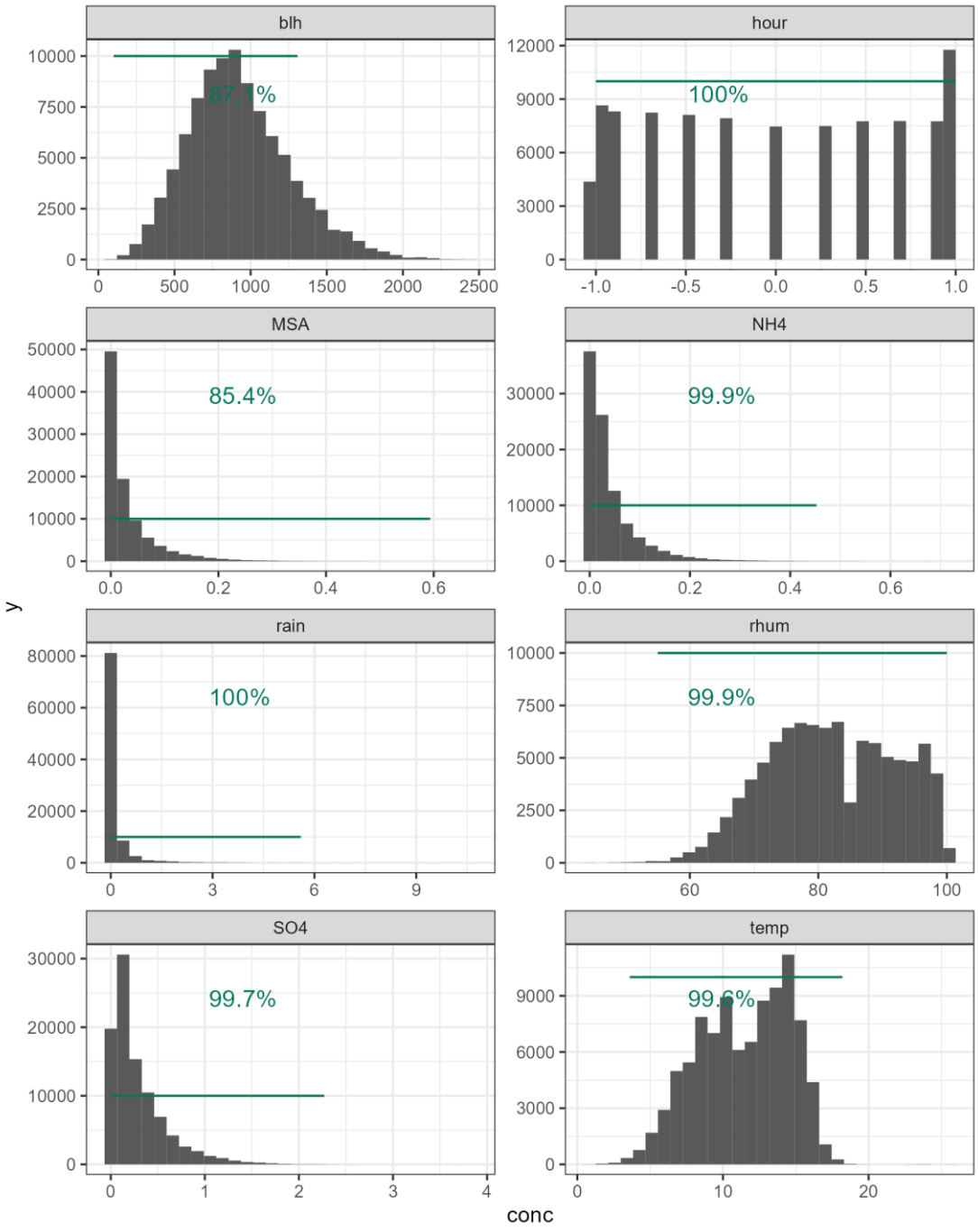


Figure S1. Frequency distribution of each predictor across the entire clean marine dataset and the range of the data used in SOA model training, the label indicates the data coverage. The hour of the day to periodic functions is simulated with cosine functions hour = cos(hour*(2pi/12)).

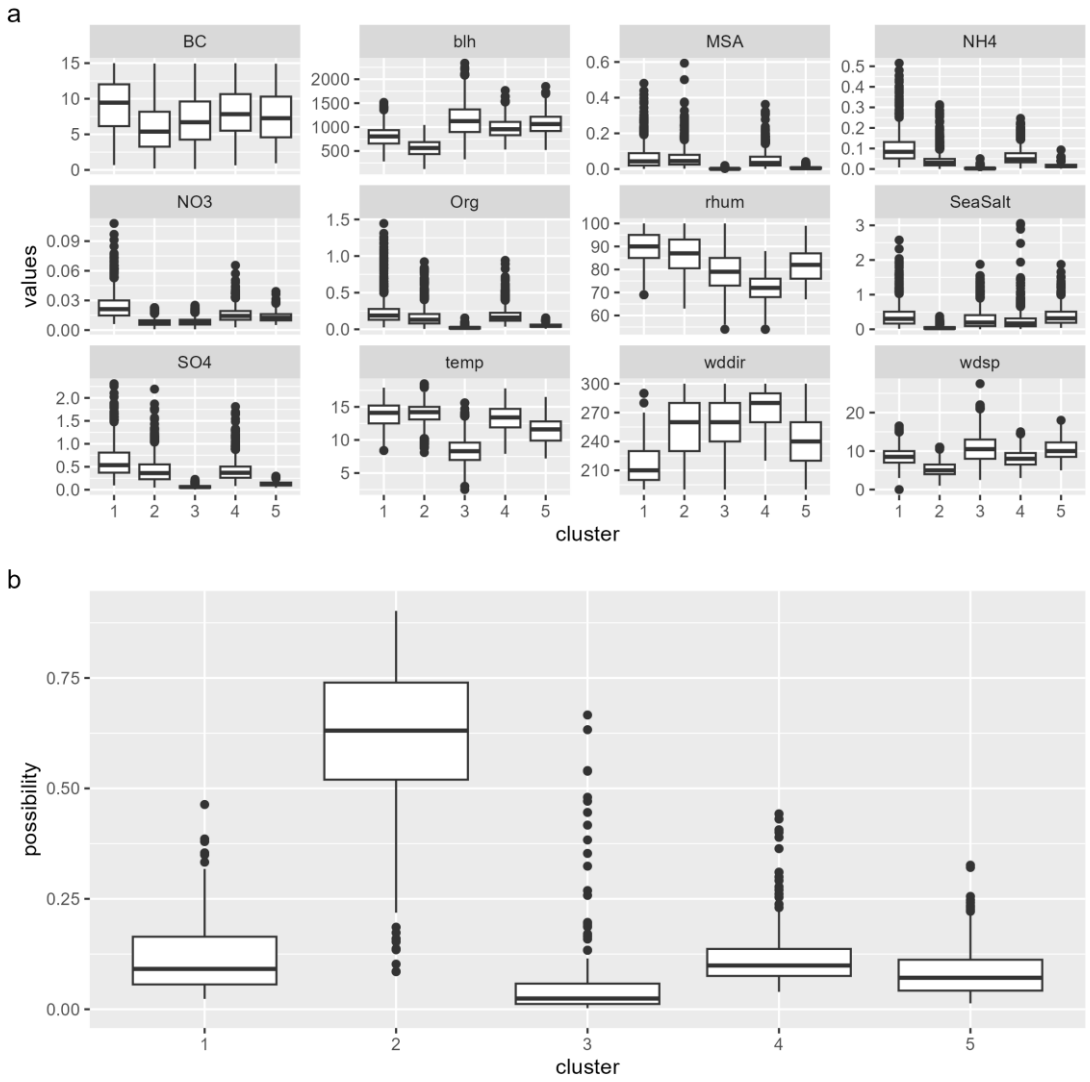


Figure S2. (a) Factor Profile Resolution via Fuzzy C-Means Clustering: This figure displays the profiles of factors resolved using fuzzy c-means clustering. To enhance clarity and relevance, only data with a cluster membership confidence exceeding 80% are included. This visualization helps in identifying 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.

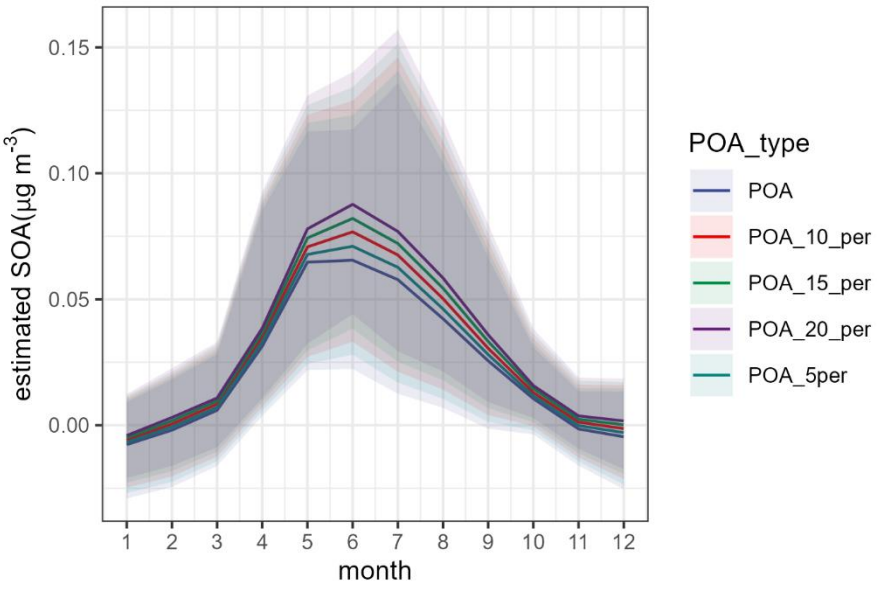


Figure S3. POA seasonality with different sensitivity test. Assuming the POA contribution (5%, 10%, 15% and 20%) to the total OA in selected secondary marine aerosol data. The shaded area represents 25th to 75th percentile of each test. Assuming that POA accounts for different percentages of total OA is likely to produce many negative predictions, especially in wintertime.

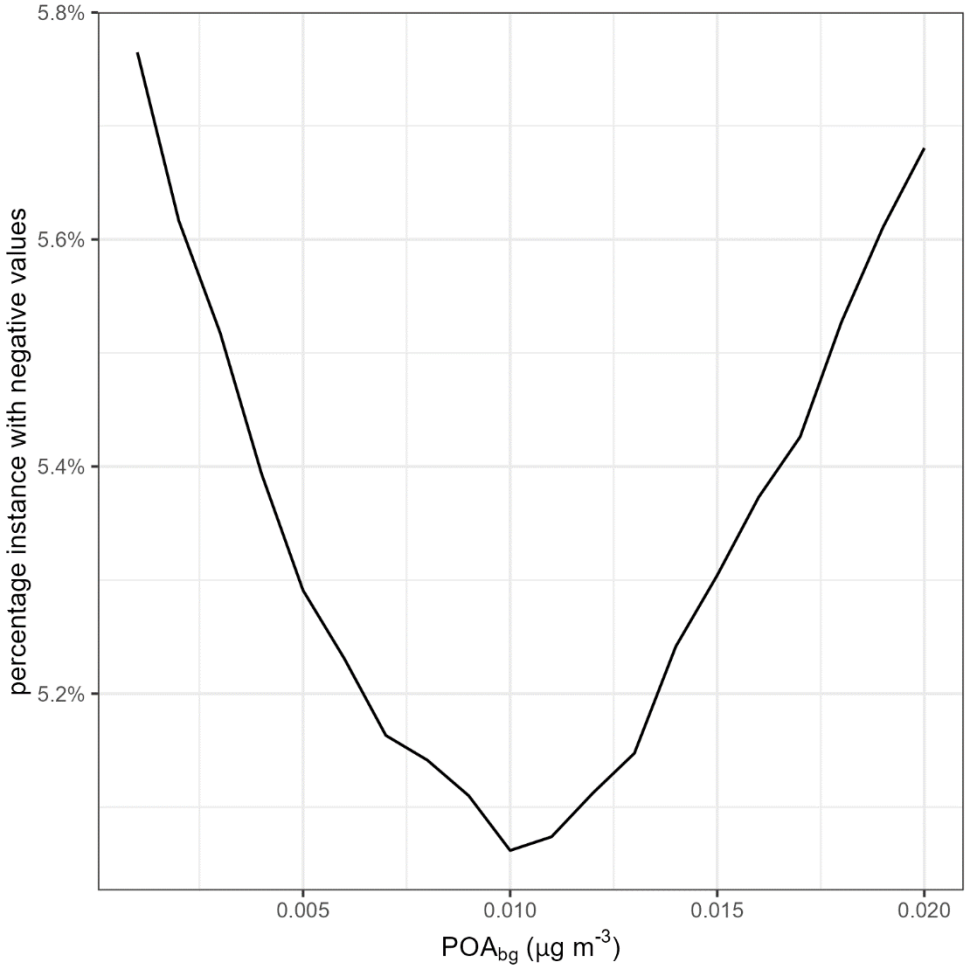


Figure S4. The different POA_{bg} value used in ML model and the percentage of non-physical prediction instance including POA and SOA (lower than 0). This indicates the POA_{bg} of 0.01 µg m^{-3} is the optimum value.

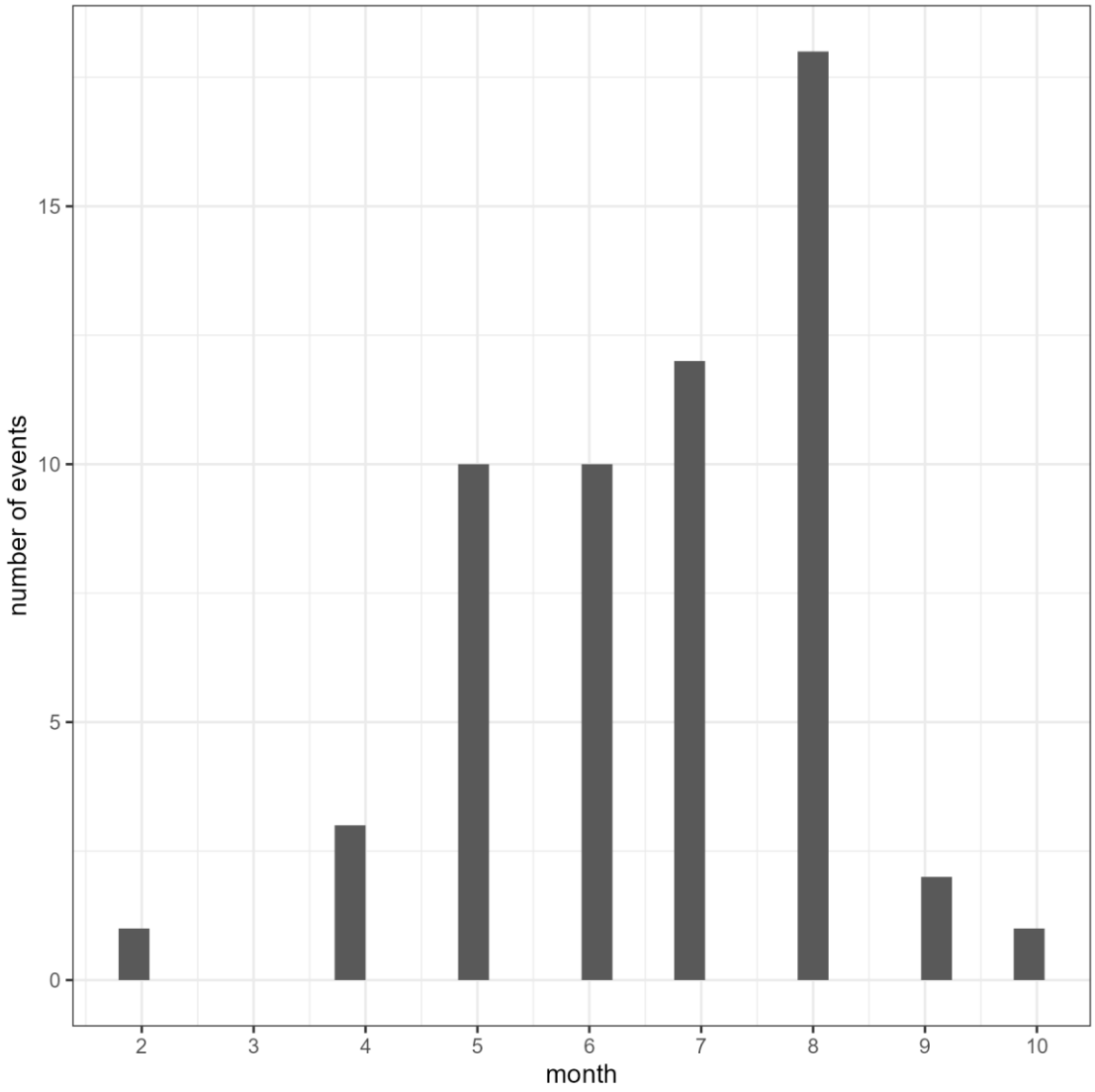


Figure S5. The monthly distribution of POA events. POA event is defined as a period with POA concentration higher than 0.1 μg m^{-3} over 12 hours.

40

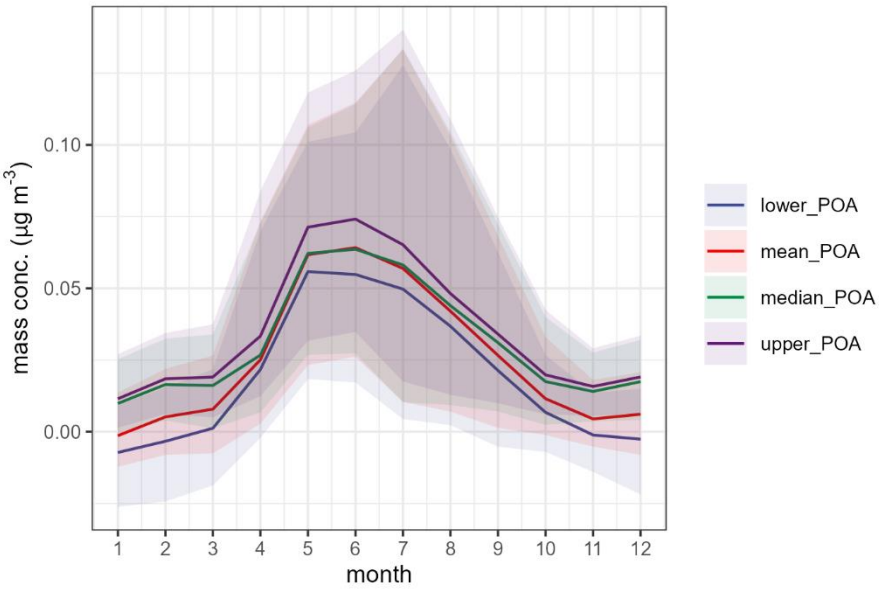


Figure S6. Monte Carlo simulations. The Monte Carlo simulation is done by randomly drop 20% of the data in training dataset and repeat the training process for 1000 times. The mean, median, lower (25th percentile) and upper (75th percentile) of the simulation ensemble were summarised monthly.

45

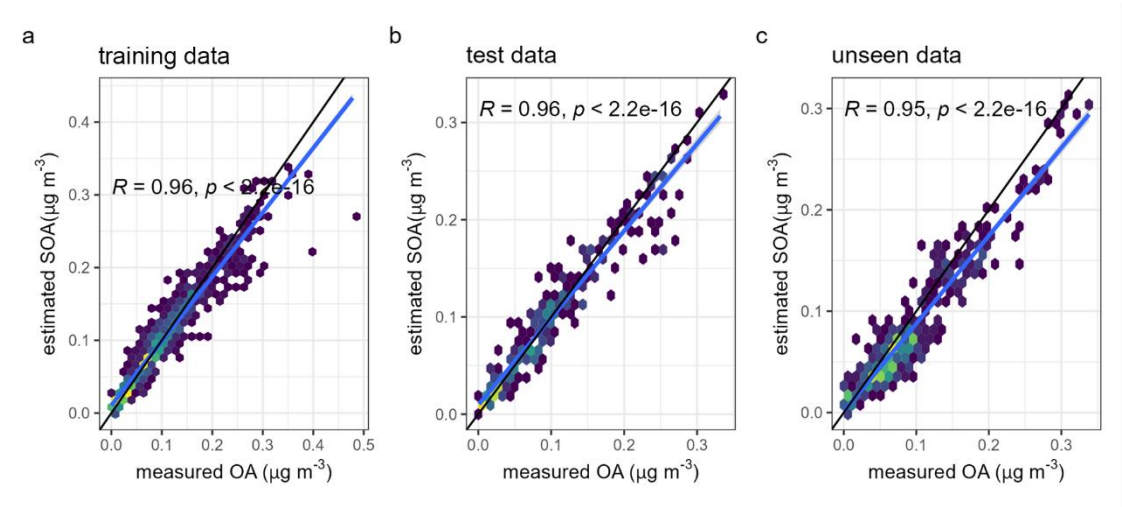


Figure S7. Observed OA versus predicted (SOA +POA_{bg}) for (a) training, (b) validation and (c) test datasets. Data density is illustrated using a color gradient with darker colour indicating lower data density. Black lines denote the 1:1 correspondence lines, blue lines represent regression lines. The ML model does not use MSA as predictors.

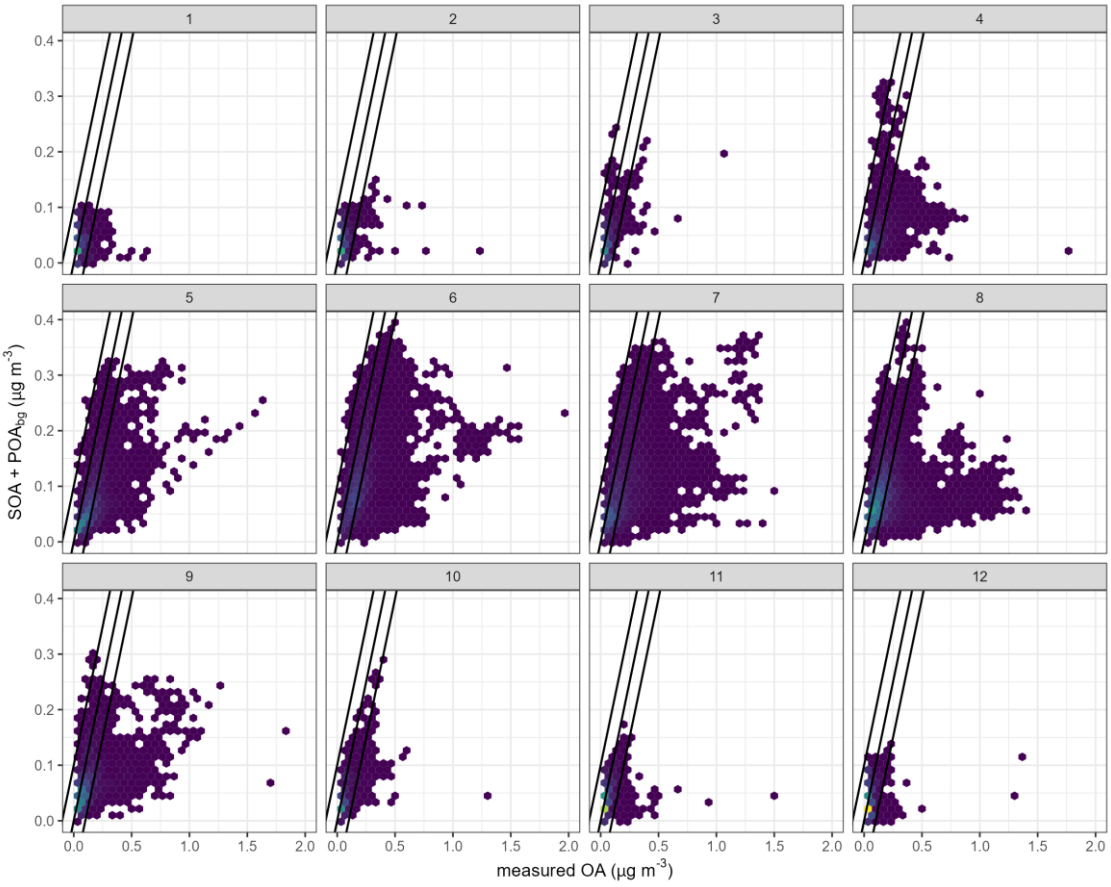
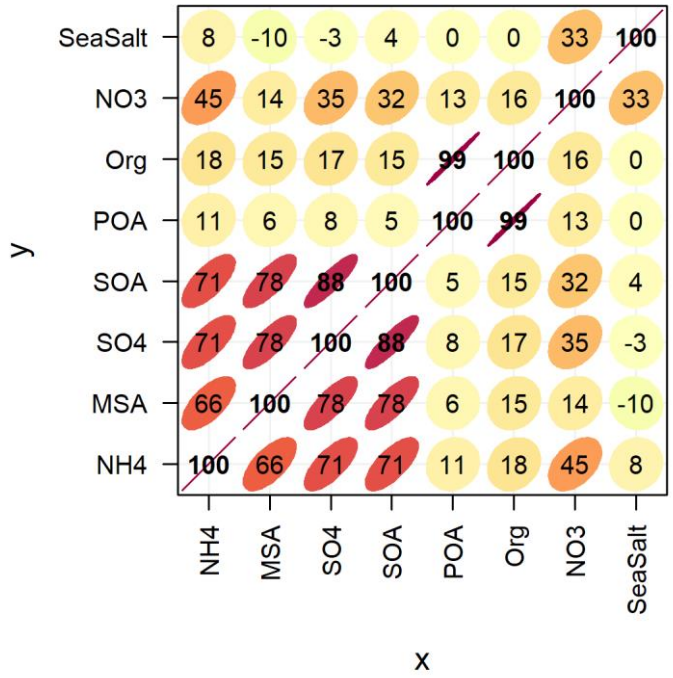


Figure S8. The estimated SOA + POA_{bg} versus measured total OA in different months. The black lines represent y = x - 0.1, y = x, and y = x+0.1.



55

Figure S9. Correlation between chemical species for the entire clean marine air mass data.