S1 Fuzzy clustering validity indices

Six fuzzy validity indices were used to determine the appropriate number of clusters, include Sum of within-cluster variance (V_{SWCV}), Fakuyama-Sugeno index (V_{FS}, Fukuyama, 1989), Xie-Beni index (V_{XB}, Xie and Beni, 1991), Kwon index (V_{Kwon}, Kwon, 1998), Bouguessa-Wang-Sun index (V_{BWS}, Bouguessa et al., 2006), and Fuzzy Silhouette (FS, Campello and Hruschka, 2006). Their definitions and notes for applications are described in this section.

(1) Sum of within-cluster variation (V_{SWCV}). The basic idea of clustering is to sort clusters so that the sum of within-cluster variation is minimized, and this is used as the objective function $J_m(U,V)$ in fuzzy *c*-means clustering, as given by Eq. S1. The sum of within-cluster squared distance measures the compactness of clustering, and the "knee" in the curve of V_{SWCV} as a function of numbers of clusters is generally considered as an indicator of the optimal number of clusters (Campello and Hruschka, 2006).

$$V_{SWCV} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|x_j - v_i\right\|^2 \quad \text{(S1)}$$

where x_j and v_i denote the j^{th} object in the dataset and the i^{th} cluster center, respectively, m is the fuzzifier, and u_{ij} is the membership degree of x_j to the i^{th} cluster.

(2) Fukuyama-Sugeno index (V_{FS}). The Fakuyama-Sugeno index combines the membership degree and the geometrical property of the dataset to evaluate a partition (Bouguessa and Wang, 2004). It evaluates the quality of a clustering solution by measuring the discrepancy between compactness and separation of clusters. The mathematical expression of V_{FS} is shown in Eq. S2, where the sum of within-cluster variances, as the first item in the equation, represents compactness, while the sum of squared distances between each cluster center and the mean of all cluster centers, as the second item in the equation, measures the separation pf partition. Obviously, smaller V_{FS} indicates better performance of clustering.

$$V_{FS} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|x_j - v_i\right\|^2 - \sum_{i=1}^{c}(\sum_{j=1}^{n} u_{ij}^{m})\|v_i - \bar{v}\|^2 \quad \text{(S2)}$$

where $\bar{v} = \frac{1}{c}\sum_{i=1}^{c} v_i$.

Identically, x_j and v_i denote the j^{th} object in the dataset and the i^{th} cluster center, respectively, m is the fuzzifier, and u_{ij} is the membership degree of x_j to the i^{th} cluster.

(3) Xie-Beni index (V_{XB}). Xie-Beni index is a popular fuzzy clustering validity measure proposed by Xie and Beni (1991). It is defined as the ratio of compactness and separation as shown in Eq. S3, where the sum of within-cluster squared distance which is equivalent to the objective function $J_m(U,V)$, divided by the total number of objects in the numerator, is the compactness of the partition, and the minimum squared distance of cluster centers in the denominator is represents the separation. The smaller the numerator, the more compact is a cluster, whereas the larger the denominator, the more a cluster is separated. As a consequence, the smaller V_{XB}, the better the partition.

$$V_{XB} = \frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\|x_j - v_i\|^2}{\min\limits_{k\neq i}\|v_k - v_i\|^2} \quad \text{(S3)}$$

where x_j, v_i and v_k denote the j^{th} object in the dataset, the i^{th} and k^{th} cluster center, respectively, m is the fuzzifier, and u_{ij} is the membership degree of x_j to the i^{th} cluster.

(4) Kwon index (V_{kwon}). When c approaches n, the value of V_{XB} decreases monotonically to 0 and will lose robustness in determining the optimal number of clusters. To overcome this drawback, Kwon (1998) revised V_{XB} and proposed Kwon index, as defined in Eq. S4. The second item in the numerator is a punishing function, which represents the average squared distance of cluster centers to the overall mean of the data set and can eliminate its monotonous decreasing tendency when the number of clusters is close to n. Similar to V_{XB}, the smaller V_{kwon}, the better the clustering quality.

$$V_K = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\|x_j - v_i\|^2 + (1/c)\sum_{i=1}^{c}\|v_i - \bar{x}\|^2}{\min\limits_{k\neq i}\|v_k - v_i\|^2} \quad \text{(S4)}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$.

Identically, x_j, v_i and v_k in the formula denote the j^{th} object in the dataset, the i^{th} and k^{th} cluster center, respectively, m is the fuzzifier, and u_{ij} is the membership degree of x_j to the i^{th} cluster.

(5) Bouguessa-Wang-Sun index (V_{BWS}). To better deal with overlapped clusters that differ in geometric shape, Bouguessa et al. (2006) proposed a new validity index, as formulated in Eq. S5, and hereafter called Bouguessa-Wang-Sun index in this study. Similar to V_{XB} and V_{Kwon}, V_{BWS} is also based on the concept of using the ratio of compactness and separation, but the definitions for

compactness and separation are modified. By making use of the fuzzy covariance matrix as a measure of compactness, V_{BWS} takes the variations of cluster shape, density and orientation into account and was proved to performe well for heavily overlapping clusters (Bouguessa and Wang, 2004; Bouguessa et al., 2006). According to its definition, a larger value of V_{BWS} indicates a better fuzzy partition.

$$V_{BWS} = \frac{Sep(c)}{Comp(c)} \tag{S5}$$

In the equation, $Sep(c)$ represents fuzzy separation, as defined in Eq. S6, and S_B is the between-cluster fuzzy matrix given by Eq. S7. The larger $Sep(c)$, the better separation between clusters.

$$Sep(c) = trace(S_B) \tag{S6}$$

$$S_B = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} (v_i - \bar{v}) (v_i - \bar{v})^T \tag{S7}$$

$Comp(c)$ in Eq. S5 represents the overall compactness of fuzzy clustering, as given by Eq. S8. The smaller $Comp(c)$, the more compact within each cluster.

$$Comp(c) = \sum_{i=1}^{c} trace(\sum_i) \tag{S8}$$

where $\sum_i$ is the fuzzy covariance matrix as defined by:

$$\sum_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{S9}$$

(6) Fuzzy Silhouette (FS). The silhouette score (s_j, as defined in Eq. S10) was first proposed by Rousseeuw (1987), which can be used to measure how close an object is to the cluster center it belongs compared to other clusters. The average silhouette score of all objects, CS, as given by Eq. S11, are frequently used to assess the quality of clustering solutions. The silhouette score was originally adopted to evaluate hard or non-fuzzy clustering solutions and did not consider the fuzzy partition matrix in the calculation. Consequently, CS might be inadequate to discriminate fuzzy clusters since it ignores the information contained in the fuzzy partition matrix which reveal the overlap degrees of clusters. To extend the silhouette score to fuzzy partition and make explicit use of the fuzzy partition matrix, Campello and Hruschka (2006) proposed *Fuzzy Silhouette* (FS), as given by Eq. S12. Instead of weighing each individual silhouette equally, FS stresses the importance of objects lying in the vicinity of cluster centers

while reducing the importance of objects located in the boundary region (whose membership degrees to different clusters are similar or identical).

The silhouette score falls in the range from -1 to +1, with a value approaching +1 indicating that the object is correctly assigned, whereas with a value close to -1 indicating that the object is misclustered (better to sort it to a neighboring cluster than to current cluster). An s_j close to 0 implies that the object lies in the boundary region (between clusters) and thus it's unclear to which cluster it belongs. The average cluster silhouette score can tell if the cluster is appropriately configurated or not. The larger the average cluster silhouette score, the clearer the cluster. The overall average silhouette score of all objects in the dataset can be used as a measure of clustering quality. Further, it can be used to find the appropriate number of clusters. With different cluster numbers, the maximum overall silhouette score, which means minimizing the intra-cluster distance (a_{pj}) while maximizing the inter-cluster distance (b_{pj}), indicates the optimal number of clusters.

$$s_j = \frac{b_{pj} - a_{pj}}{max\{a_{pj}, b_{pj}\}} \tag{S10}$$

where a_{pj} is the average distance of object j (belonging to cluster p) to all other objects in the same cluster. Let d_{qj} be the average distance of object j to all objects belonging to another cluster q ($q \neq p$), then b_{pj} is the minimum d_{qj}, which represents the average distance of object j to its closet neighboring cluster.

$$CS = \frac{1}{n}\sum_{j=1}^{n} s_j \tag{S11}$$

$$FS = \frac{\sum_{j=1}^{n}(u_{pj} - u_{qj})^{\alpha} s_j}{\sum_{j=1}^{n}(u_{pj} - u_{qj})^{\alpha}} \tag{S12}$$

where s_j in the average silhouette score of object j calculated according to Eq. S10, u_{pj} and u_{qj} are the first and second largest coefficient in column j of the fuzzy partition matrix, respectively, and α is a weight coefficient and set to be 1 as default in this study (Campello and Hruschka, 2006).

Table S1. Possible permutation scheme for 2N- (grey), 3N- (blue) and 4N-dimers (orange) formed through $RO_2 + R'O_2$ reactions. Second-generation species are outlined in blue.

$C_5H_8NO_x$ \ $C_5H_8NO_x$	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}
O_5	$C_{10}H_{16}N_2O_8$						
O_6	$C_{10}H_{16}N_2O_9$	$C_{10}H_{16}N_2O_{10}$					
O_7	$C_{10}H_{16}N_2O_{10}$	$C_{10}H_{16}N_2O_{11}$	$C_{10}H_{16}N_2O_{12}$				
O_8	$C_{10}H_{16}N_2O_{11}$	$C_{10}H_{16}N_2O_{12}$	$C_{10}H_{16}N_2O_{13}$	$C_{10}H_{16}N_2O_{14}$			
O_9	$C_{10}H_{16}N_2O_{12}$	$C_{10}H_{16}N_2O_{13}$	$C_{10}H_{16}N_2O_{14}$	$C_{10}H_{16}N_2O_{15}$	$C_{10}H_{16}N_2O_{16}$		
O_{10}	$C_{10}H_{16}N_2O_{13}$	$C_{10}H_{16}N_2O_{14}$	$C_{10}H_{16}N_2O_{15}$	$C_{10}H_{16}N_2O_{16}$	$C_{10}H_{16}N_2O_{17}$	$C_{10}H_{16}N_2O_{18}$	
O_{11}	$C_{10}H_{16}N_2O_{14}$	$C_{10}H_{16}N_2O_{15}$	$C_{10}H_{16}N_2O_{16}$	$C_{10}H_{16}N_2O_{17}$	$C_{10}H_{16}N_2O_{18}$	$C_{10}H_{16}N_2O_{19}$	$C_{10}H_{16}N_2O_{20}$
$C_5H_8NO_x$ \ $C_5H_9N_2O_y$	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}
O_9	$C_{10}H_{17}N_3O_{12}$	$C_{10}H_{17}N_3O_{13}$	$C_{10}H_{17}N_3O_{14}$	$C_{10}H_{17}N_3O_{15}$	$C_{10}H_{17}N_3O_{16}$	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$
O_{10}	$C_{10}H_{17}N_3O_{13}$	$C_{10}H_{17}N_3O_{14}$	$C_{10}H_{17}N_3O_{15}$	$C_{10}H_{17}N_3O_{16}$	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$
O_{11}	$C_{10}H_{17}N_3O_{14}$	$C_{10}H_{17}N_3O_{15}$	$C_{10}H_{17}N_3O_{16}$	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$
O_{12}	$C_{10}H_{17}N_3O_{15}$	$C_{10}H_{17}N_3O_{16}$	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$	$C_{10}H_{17}N_3O_{21}$
O_{13}	$C_{10}H_{17}N_3O_{16}$	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$	$C_{10}H_{17}N_3O_{21}$	$C_{10}H_{17}N_3O_{22}$
O_{14}	$C_{10}H_{17}N_3O_{17}$	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$	$C_{10}H_{17}N_3O_{21}$	$C_{10}H_{17}N_3O_{22}$	$C_{10}H_{17}N_3O_{23}$
O_{15}	$C_{10}H_{17}N_3O_{18}$	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$	$C_{10}H_{17}N_3O_{21}$	$C_{10}H_{17}N_3O_{22}$	$C_{10}H_{17}N_3O_{23}$	$C_{10}H_{17}N_3O_{24}$
O_{16}	$C_{10}H_{17}N_3O_{19}$	$C_{10}H_{17}N_3O_{20}$	$C_{10}H_{17}N_3O_{21}$	$C_{10}H_{17}N_3O_{22}$	$C_{10}H_{17}N_3O_{23}$	$C_{10}H_{17}N_3O_{24}$	$C_{10}H_{17}N_3O_{25}$
$C_5H_9N_2O_y$ \ $C_5H_9N_2O_y$	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}
O_9	$C_{10}H_{18}N_4O_{16}$						
O_{10}	$C_{10}H_{18}N_4O_{17}$	$C_{10}H_{18}N_4O_{18}$					
O_{11}	$C_{10}H_{18}N_4O_{18}$	$C_{10}H_{18}N_4O_{19}$	$C_{10}H_{18}N_4O_{20}$				
O_{12}	$C_{10}H_{18}N_4O_{19}$	$C_{10}H_{18}N_4O_{20}$	$C_{10}H_{18}N_4O_{21}$	$C_{10}H_{18}N_4O_{22}$			
O_{13}	$C_{10}H_{18}N_4O_{20}$	$C_{10}H_{18}N_4O_{21}$	$C_{10}H_{18}N_4O_{22}$	$C_{10}H_{18}N_4O_{23}$	$C_{10}H_{18}N_4O_{24}$		
O_{14}	$C_{10}H_{18}N_4O_{21}$	$C_{10}H_{18}N_4O_{22}$	$C_{10}H_{18}N_4O_{23}$	$C_{10}H_{18}N_4O_{24}$	$C_{10}H_{18}N_4O_{25}$	$C_{10}H_{18}N_4O_{26}$	
O_{15}	$C_{10}H_{18}N_4O_{22}$	$C_{10}H_{18}N_4O_{23}$	$C_{10}H_{18}N_4O_{24}$	$C_{10}H_{18}N_4O_{25}$	$C_{10}H_{18}N_4O_{26}$	$C_{10}H_{18}N_4O_{27}$	$C_{10}H_{18}N_4O_{28}$
O_{16}	$C_{10}H_{18}N_4O_{23}$	$C_{10}H_{18}N_4O_{24}$	$C_{10}H_{18}N_4O_{25}$	$C_{10}H_{18}N_4O_{26}$	$C_{10}H_{18}N_4O_{27}$	$C_{10}H_{18}N_4O_{28}$	$C_{10}H_{18}N_4O_{29}$

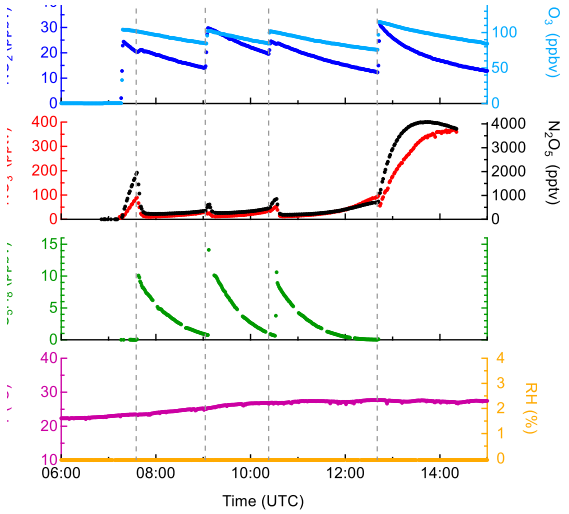


Figure S1. Concentrations of trace gases (NO_x, NO_y, and isoprene) and conditions of the chamber experiment selected for FCM analysis in this study. Adapted from Wu et al. (2021).

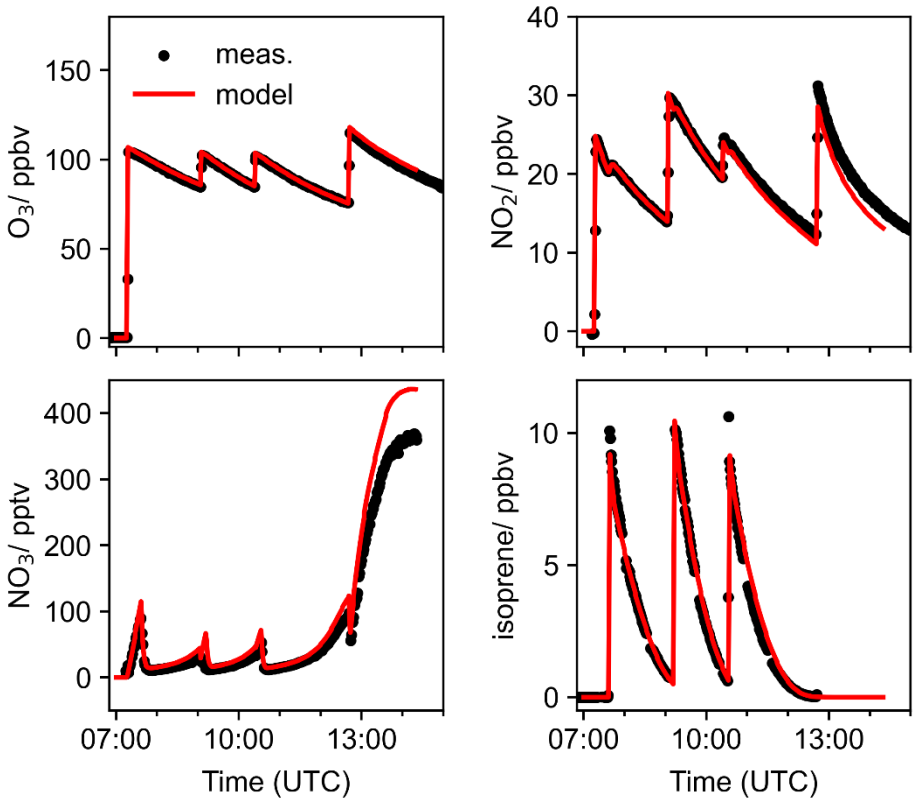


Figure S2. Measured and simulated concentrations of O3, NO2, NO3, and isoprene in the chamber experiment of isoprene oxidation by NO_3. Simulation results are from a box model with using the gas-phase chemistry mechanism of isoprene + NO3 from MCM v3.3.1.

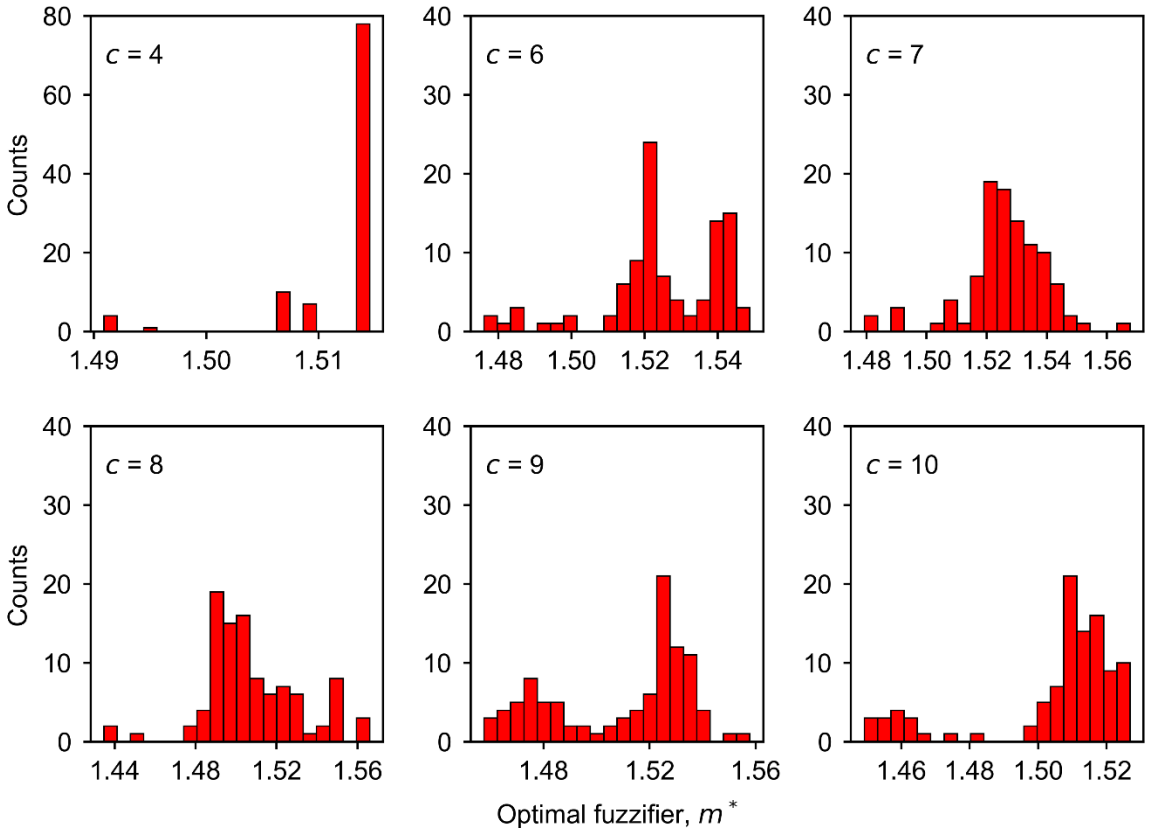
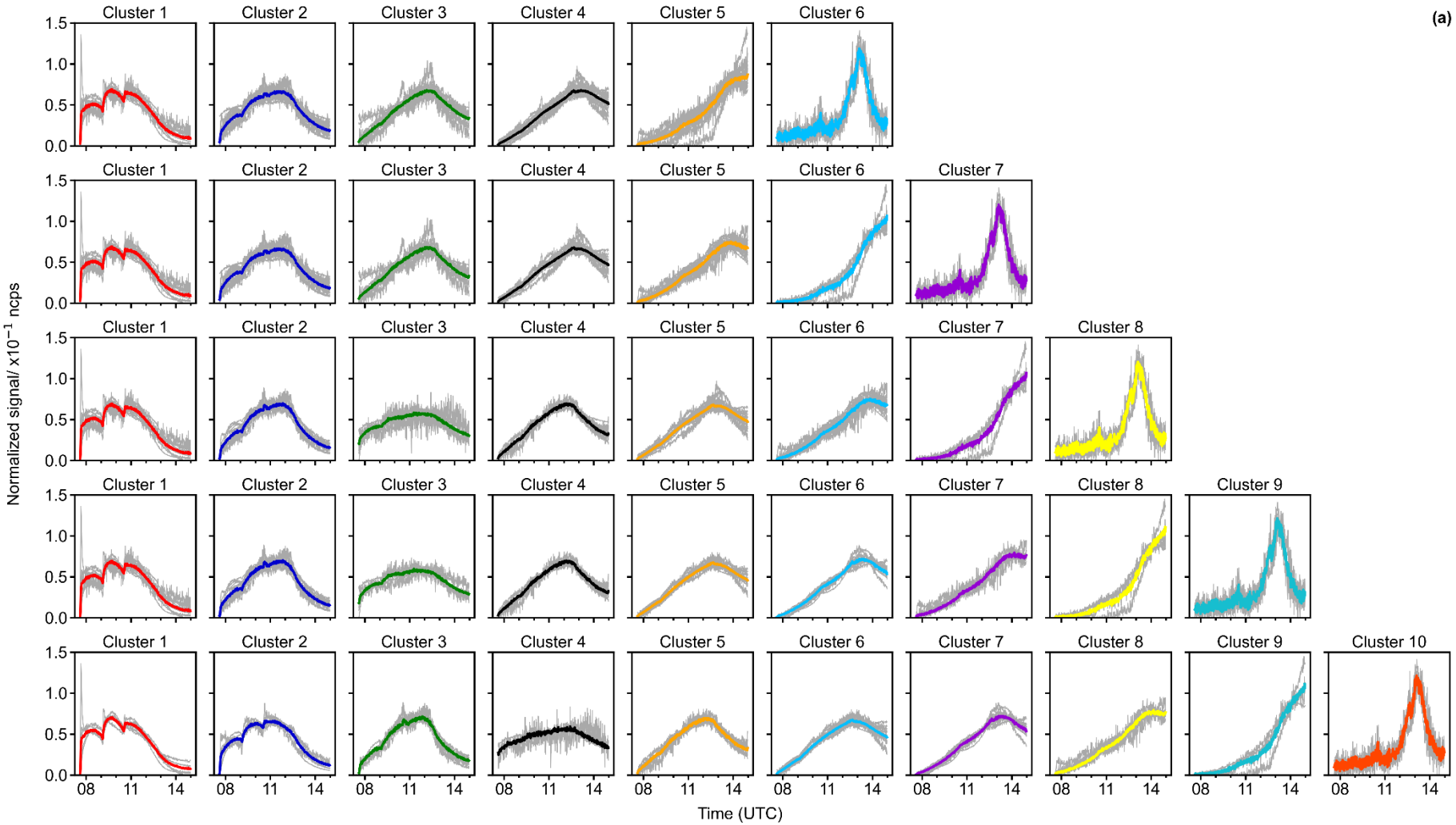


Figure S3. Distribution of the optimal value of fuzzifier (m^***)** obtained from 100 repetitions


(a)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Normalized signal/ x10^{-1} ncps
0.0
0.5
1.0
1.5
08
11
14
Time (UTC)

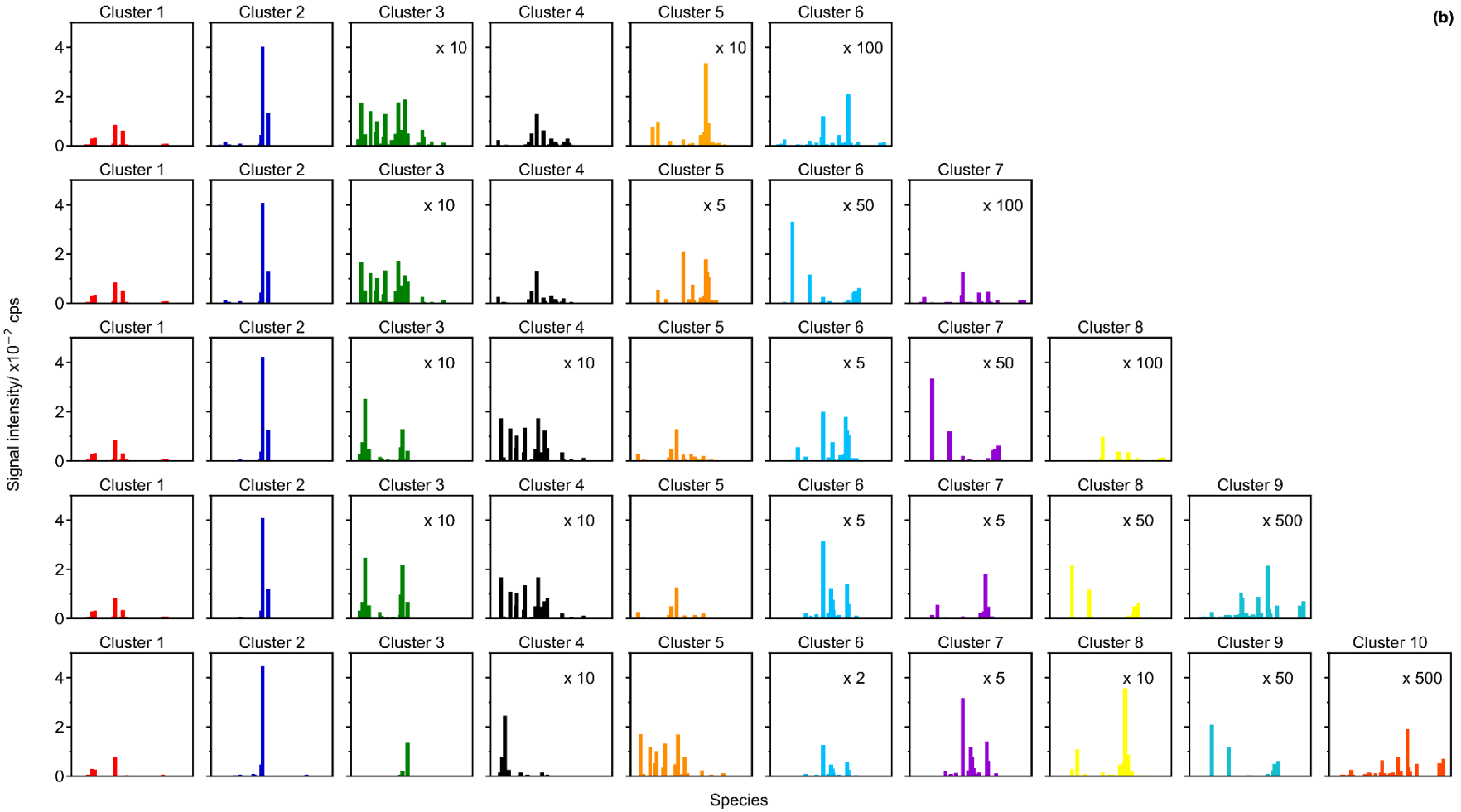


Figure S4. Fuzzy c-means clustering results of chamber data with 7-10 clusters. Time series (a) and profiles (b) of clusters for each solution. The cluster centers are shown as colored thick lines, and species with the membership degree larger than 0.5 to the cluster are illustrated as thin lines in gray.

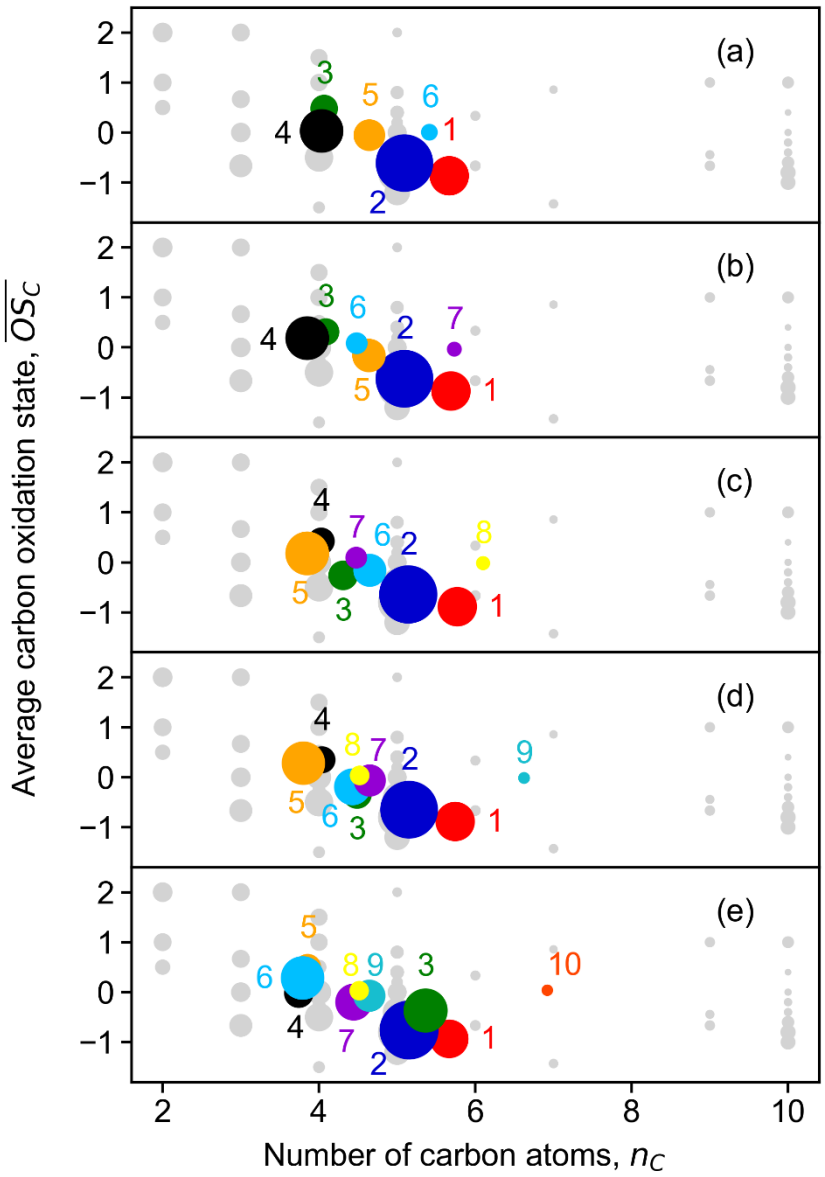


Figure S5. Average oxidation state ($\overline{OS_C}$) of FCM clusters of chamber data as a function of number of carbon atoms (n_C). Panel (a) to panel (e) show results for solutions with 6 to 10 clusters, respectively. The color scheme follows that in Fig. 4. Individual species are shown as grey circles. Marker size is proportional to the square root of the average intensity of clusters.

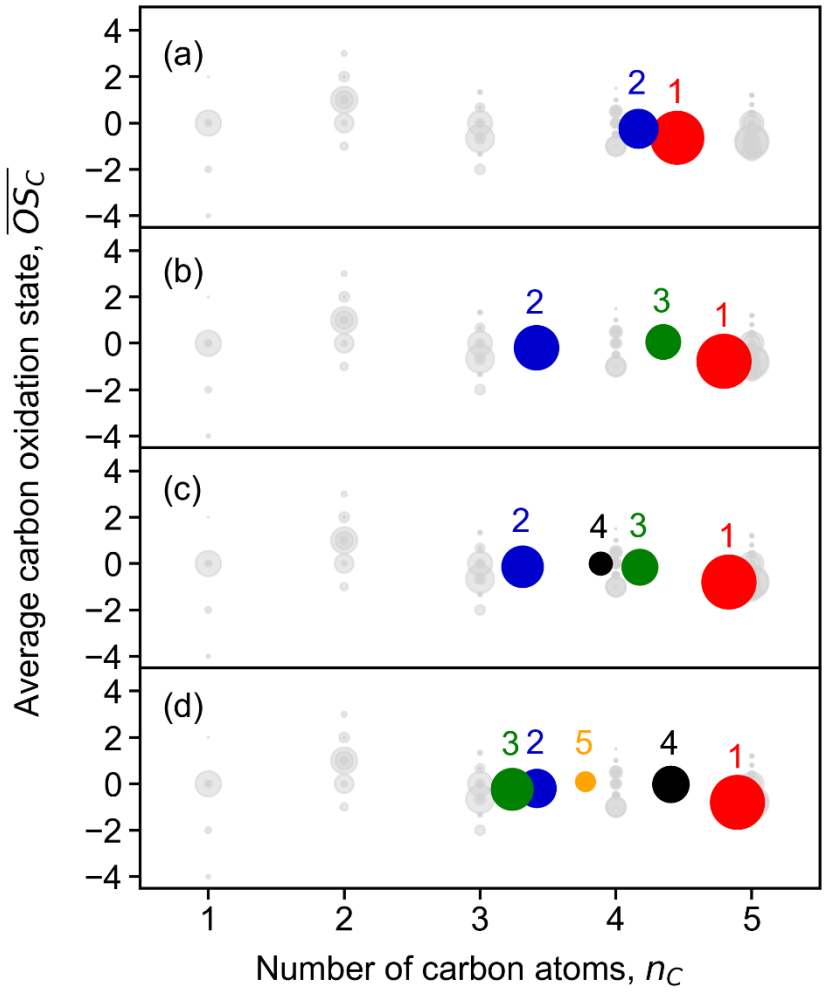


Figure S6. Average oxidation state ($\overline{OS_C}$) of FCM clusters of model data as a function of number of carbon atoms (n_C). Panel (a) to panel (d) show results for solutions with 2 to 5 clusters, respectively. The color scheme follows that in Fig. 4. Individual species are shown as grey circles. Marker size is proportional to the square root of the average intensity of clusters.

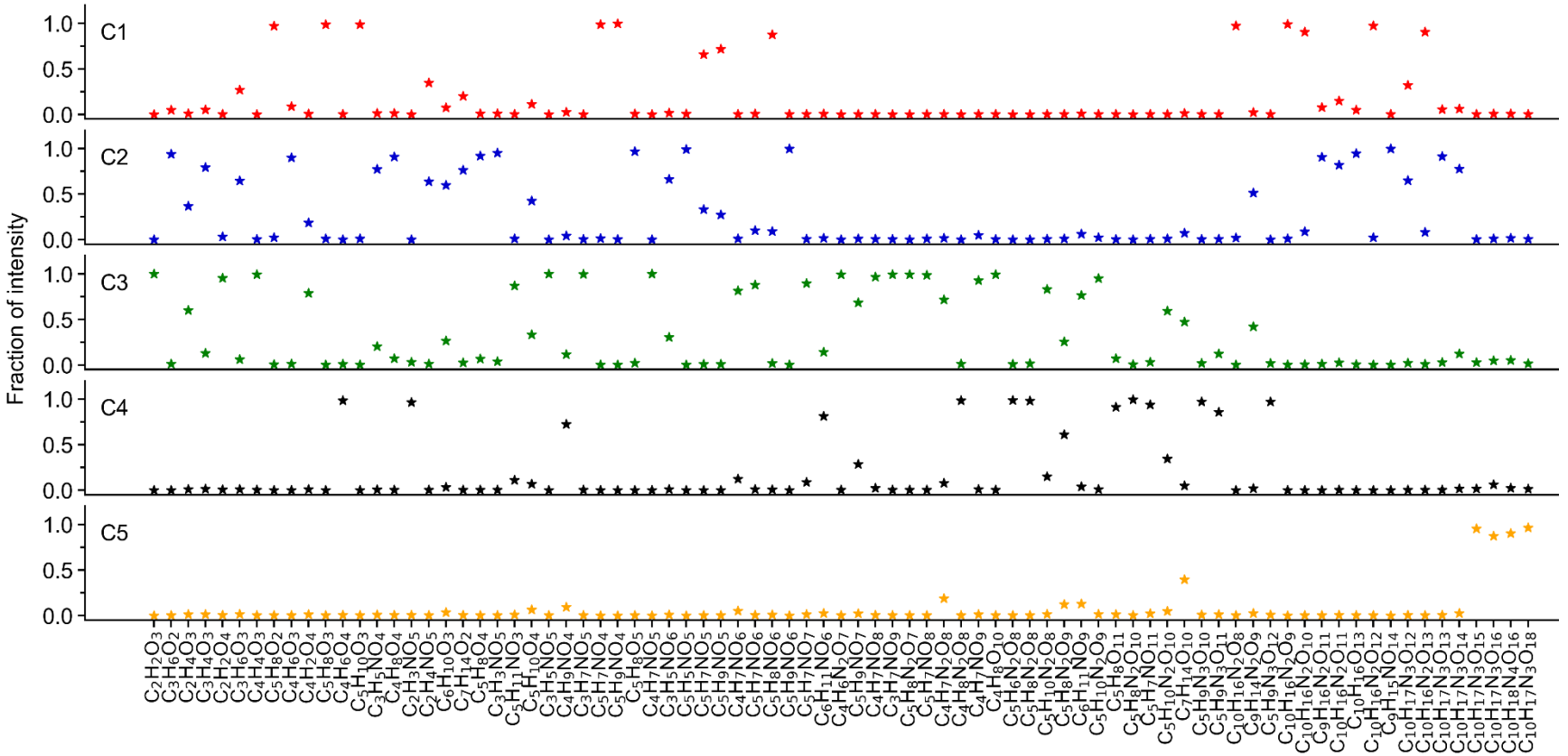


Figure S7. Cluster apportionment of species for the five-cluster solution. Sum of fractions of a compound in each cluster adds up to 1. Different clusters are distinguished by color, and the color scheme follows that in Fig. 4. Species are listed in the same order to those in Fig. 7.

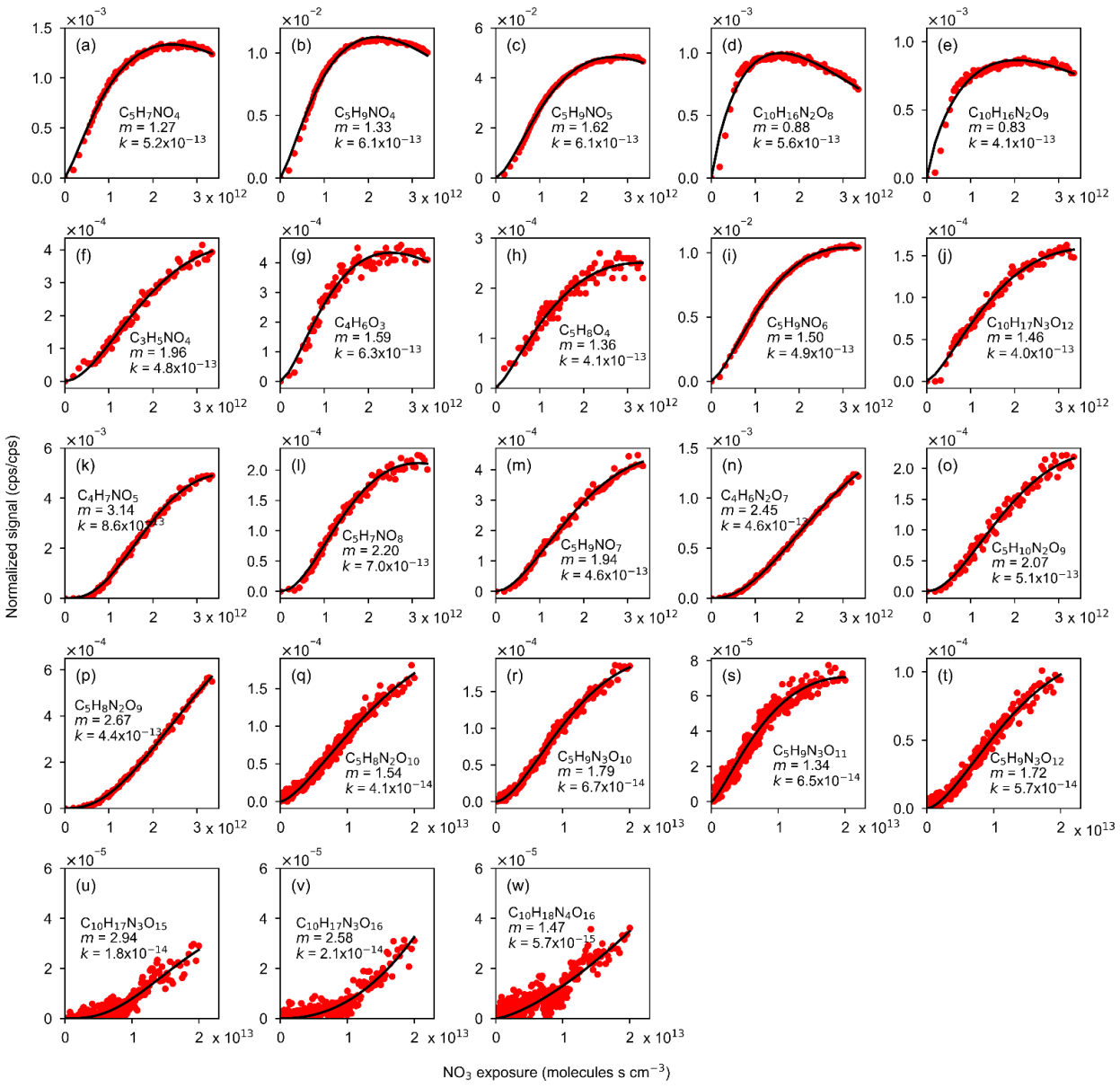
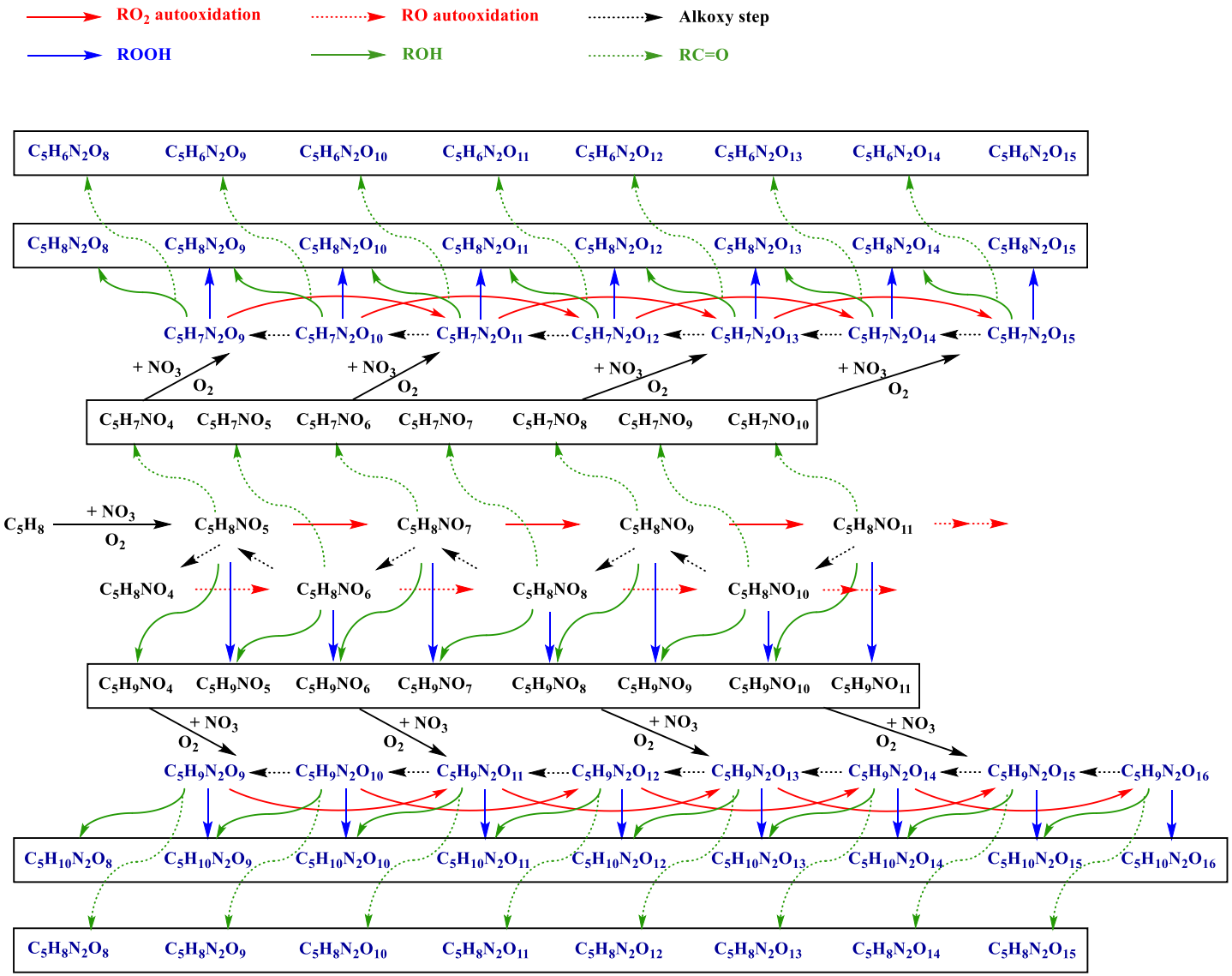
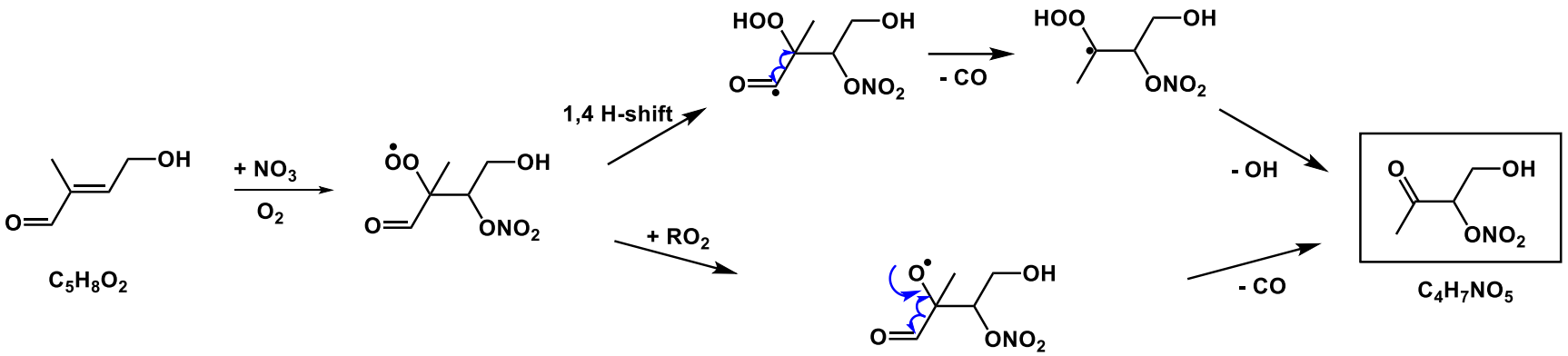


Figure S8. Representative species measured by Br^--CIMS from isoprene + NO3 experiment (red) and the GKP fitting results (black).



Scheme S1. General reaction scheme of isoprene oxidation by NO_3. The first- and second-generation products are shown in black and blue, respectively. Closed-shell species are outlined in black boxes. Dimers are not shown in this scheme for simplicity.



Scheme S2. Proposed formation mechanism of $C_4H_7NO_5$ through further oxidation of the first-generation C_5 carbonyl compound. Adapted from Wu et al. (2021).

References

Bouguessa, M. and Wang, S.-R.: A new efficient validity index for fuzzy clustering, Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826), 1914-1919,

Bouguessa, M., Wang, S., and Sun, H.: An objective approach to cluster validation, Pattern Recognition Letters, 27, 1419-1430, 10.1016/j.patrec.2006.01.015, 2006.

Campello, R. J. G. B. and Hruschka, E. R.: A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets and Systems, 157, 2858-2875, 10.1016/j.fss.2006.07.006, 2006.

Fukuyama, Y.: A new method of choosing the number of clusters for the fuzzy c-mean method, Proc. 5th Fuzzy Syst. Symp., 1989, 247-250.

Kwon, S.-H.: Cluster validity index for fuzzy clustering, Electronics Letters, 34, 2176-2177, 1998.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20, 53-65, 1987.

Xie, X. L. and Beni, G.: A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis & Machine Intelligence, 13, 841-847, 1991.