atmosphere

Supplementary Materials:

Clustering analysis on drivers of O_3 diurnal pattern and interactions with nighttime NO_3 and HONO

Xue Wang [1], Shanshan Wang [1,2,*], Sanbao Zhang [1], Chuanqi Gu [1], Aimon Tanvir [1], Ruifeng Zhang [1] and Bin Zhou [1,2,3,*]

1. DOAS fit settings and detection limits for O_3, NO_2, HCHO, HONO and NO_3

Table S1. DOAS fit settings and detection limits for O_3, NO_2, HCHO, HONO and NO_3.

Species	O_3	NO_2	HCHO	HONO	NO_3
Interferences	NO_2, HCHO, SO_2 and solar spectra	HONO, HCHO and solar spectra	NO_2, HONO, O_3 and solar spectra	HONO, HCHO and solar spectra	NO_2, H_2O and solar spectra
Fit intervals	278-290 nm	334-359 nm	335-359 nm	334-359 nm	618-669 nm
Detection limits	2.0 ppbv	1.0 ppbv	0.1 ppbv	0.1 ppbv	4 pptv

2. Validation of DOAS observations

We have compared the results of spectral retrievals with the observations of Shanghai Nanhui Environmental Monitoring Station. Since the available data is limited, we cannot compare all the results. The time series of O_3 and NO_2 obtained by the two methods in March 2020 are exhibited in Figure S1. The O_3 and NO_2 concentrations of the monitoring station were measured by Model 49i Ozone Analyzer and Model 42i NO-NO_2-NO_x Analyzer, respectively. As shown in Figure S1, both O_3 and NO_2 present highly consistent trends with the correlation coefficients R^2 more than 0.9, which testify that the DOAS fitting is credible.

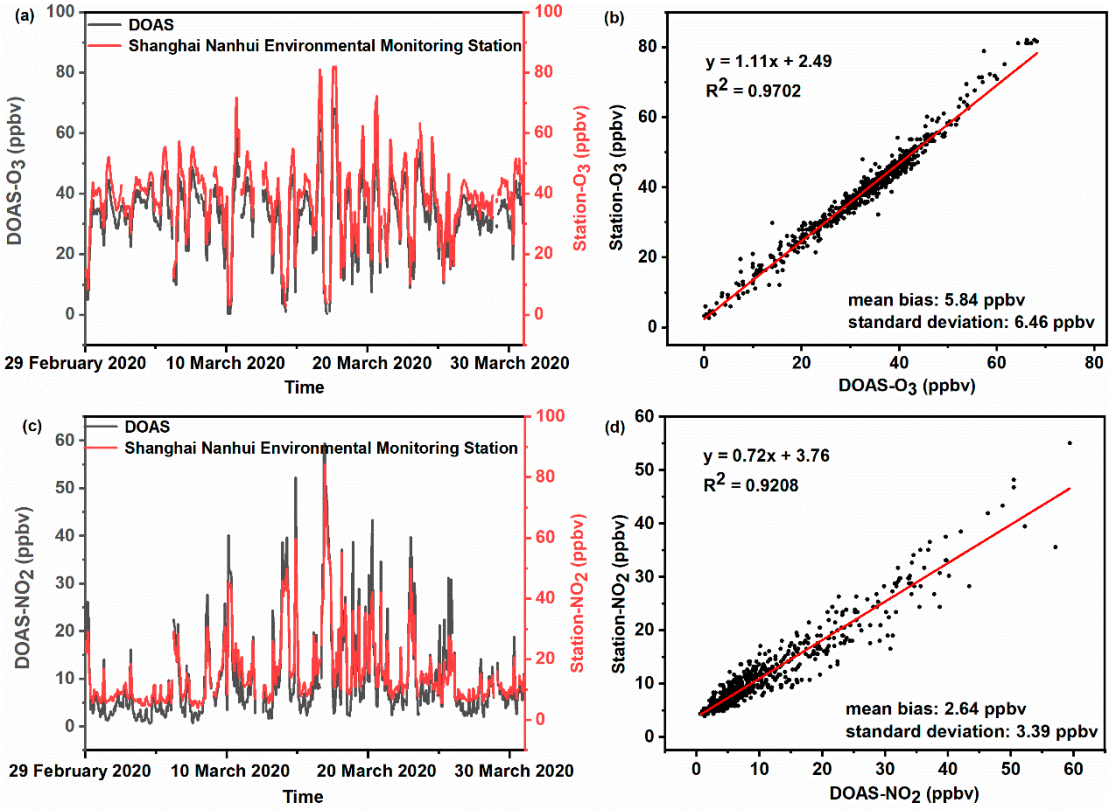

Figure S1. Time series of the O_3 (**a**) and NO_2 (**c**) mixing ratios of the DOAS retrieval and that observed by Shanghai Nanhui Environmental Monitoring Station in March 2020. And the linear fitting, mean bias and the standard deviation for the two observation methods (O_3 for (**b**) and NO_2 for Figure (**d**)).

3. Clustering analysis

- Determination of optimal cluster number

In the process of performing clustering, it is crucial for the result to choose an optimal number of clustering [1]. Although the choice is subjective, our principle is to get the most meaningful clustering result as well as the result can express an excellent effect. The Davies Bouldin index (DBI) is a prevalent indicator used to evaluate the performance of clustering [2,3]. The DBI value varies with the number of clusters. The smaller the DBI value is, the more appropriate the number is [4]. As displayed in Figure S2, we ultimately choose 4 as the number of clusters.

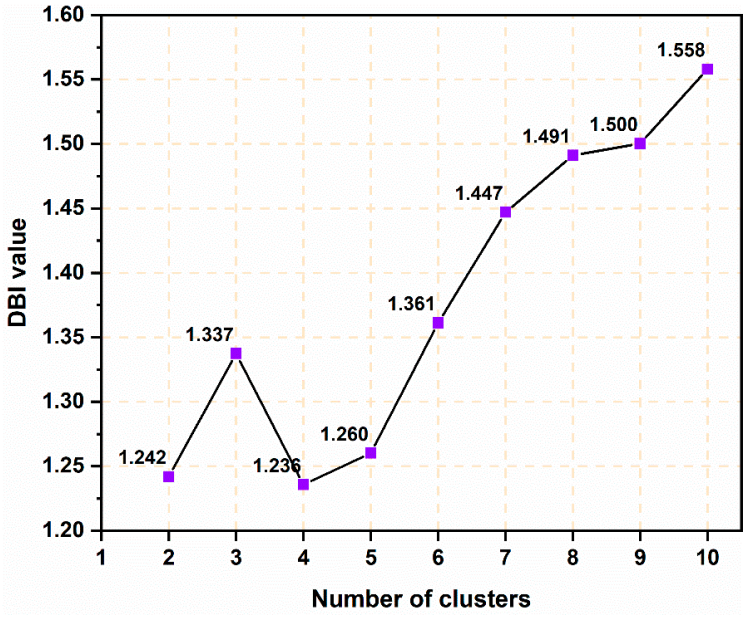

Figure S2. DBI values for the number of clusters from 1 to 10.

- Process of K-means clustering

K-means clustering is a typical clustering algorithm based on distance similarity. In this algorithm, clusters are considered to be composed of objects close to each other. The detailed process of clustering is shown in Figure S3.

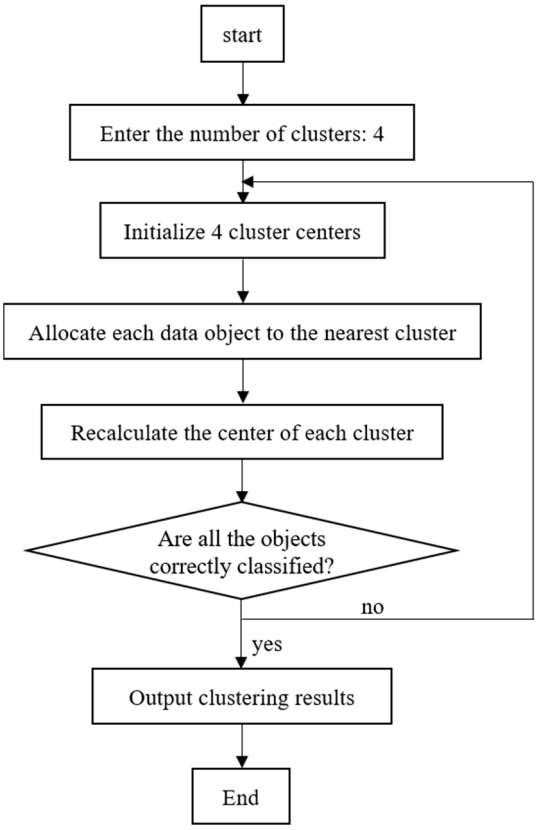

Figure S3. The process of clustering in this study.

Firstly, 4 initial cluster centers are randomly selected. The distance between all diurnal pattern objects and each cluster center is accordingly calculated, and the objects are

classified into the clusters where the nearest cluster center is located. Next, each new cluster calculates a new clustering center. Then determine whether all objects are correctly classified. If yes, the clustering result is output and the clustering process ends. If not, the next iteration starts. The symbol of end is that the cluster centers no longer change.

The visualization of clustering results is manifested in Figure 5a. Moreover, the 4 clusters used for subsequent discussion are the 4 final clustering centers.

- Evaluation of clustering results

Silhouette coefficient (SC) is a commonly used index to evaluate the clustering result [5]. SC evaluates the clustering results by combining two factors, cohesion (a) and separation (b). Cohesion refers to the dissimilarity degree within a cluster, while separation refers to the dissimilarity degree with other clusters. SC is given by

$$SC = \frac{b - a}{\max(a, b)} \quad (1)$$

Where, a is the mean distance between the target and other samples in the same cluster, and b is the mean distance between the target and all samples in other clusters. The closer SC is to 1, the more reasonable the clustering of the target is. If SC approaches to -1, the target should be divided into other clusters. When SC is close to 0, the target is on the boundary of two clusters [6]. In this study, the SC values for the 4 clusters are exhibited in Figure S4.

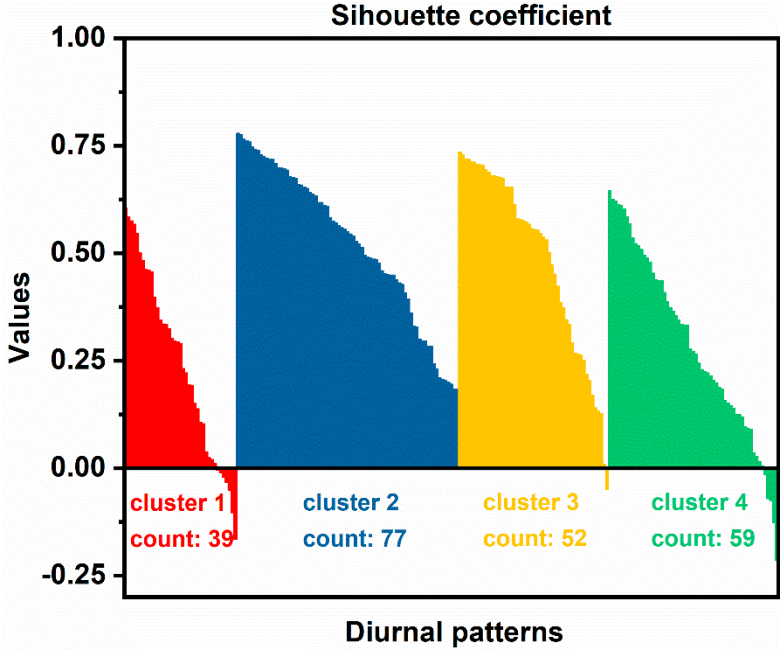

Figure S4. Bar plots of the silhouette coefficient values for all the diurnal patterns. (Diurnal patterns are sorted first by ascending cluster number and then by descending SC value.).

Almost all the SC values are more than 0, suggesting the good performance of clustering results. Compared with cluster 1 and 4, better clustering effect can be found in cluster 2 and 3. The mean SC value of all targets is 0.48. In addition, we calculated the SC values when the cluster number is defined range from 2 to 10 (Figure S5). Like the result of DBI in Figure S2, 4 is proved again to be the optimal cluster number.

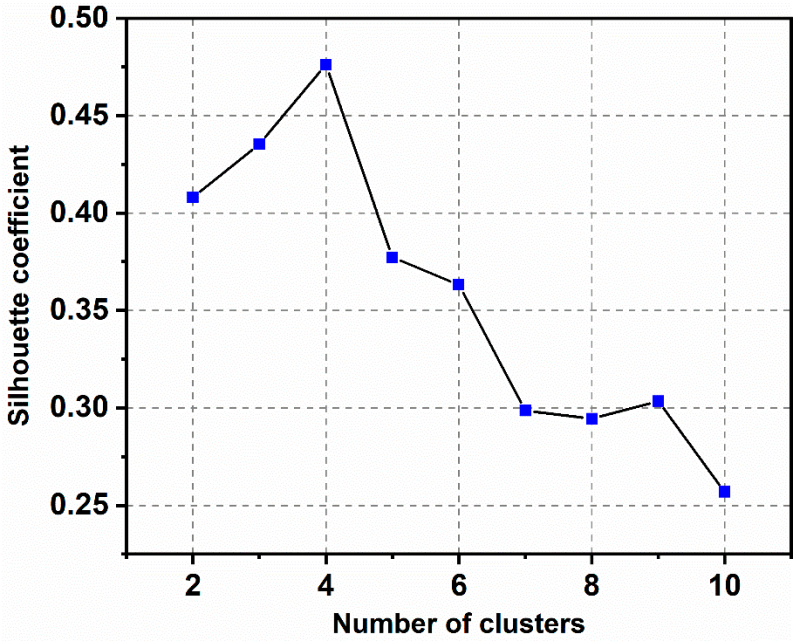

Figure S5. The SC values when the cluster number is defined range from 2 to 10.

- Monthly distribution of clusters

As can be seen in Figure S6, the total number of valid days varies from month. Cluster 1 and 4 appear more in autumn, spring and summer. And cluster 3 appears more in autumn and winter. Whereas the distribution of cluster 2 is dispersed.

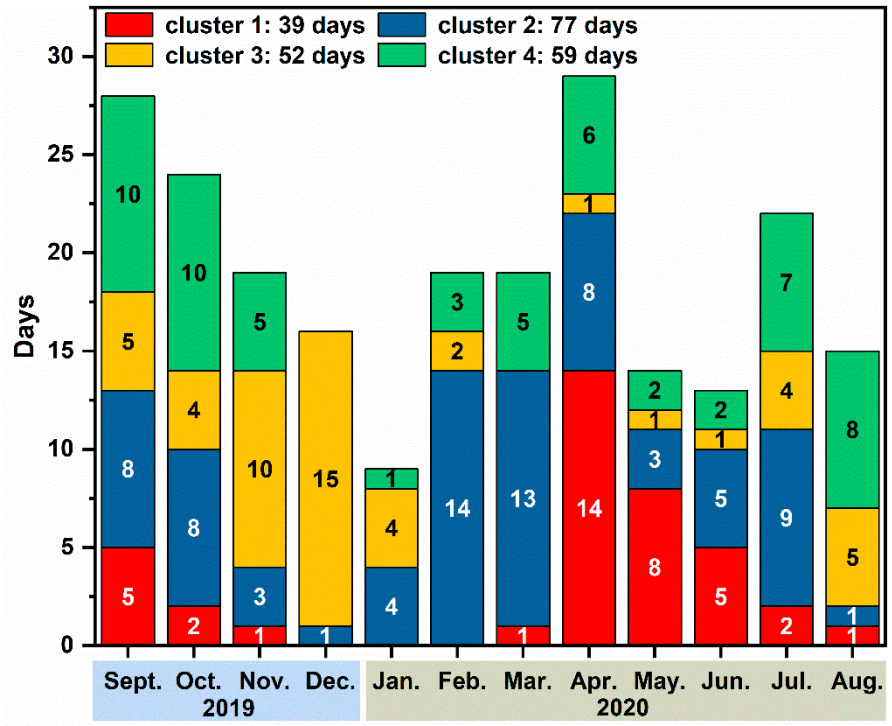

Figure S6. The monthly distribution of clustering result.

4. A case study for a high O_3 pollution episode

For studying O_3 pollution associated with the meteorological conditions and the corresponding variability of the NO_2, HCHO, HONO and NO_3 at a more detail level, an episode of four continuous days with high O_3 (April 22 to 25, 2020) is selected for further

investigation. Clusters 1 and 4 maintain higher level of O_3 increment during the daytime, so we selected an episode containing both clusters 1 and 4 for deeper analysis. The first day is from cluster 4, and the next three days are from cluster 1. The time series of atmospheric species and meteorological factors are shown in Figure S7. Moreover, we summarized the concentrations for the relevant atmospheric components at different periods as presented in Table S2.

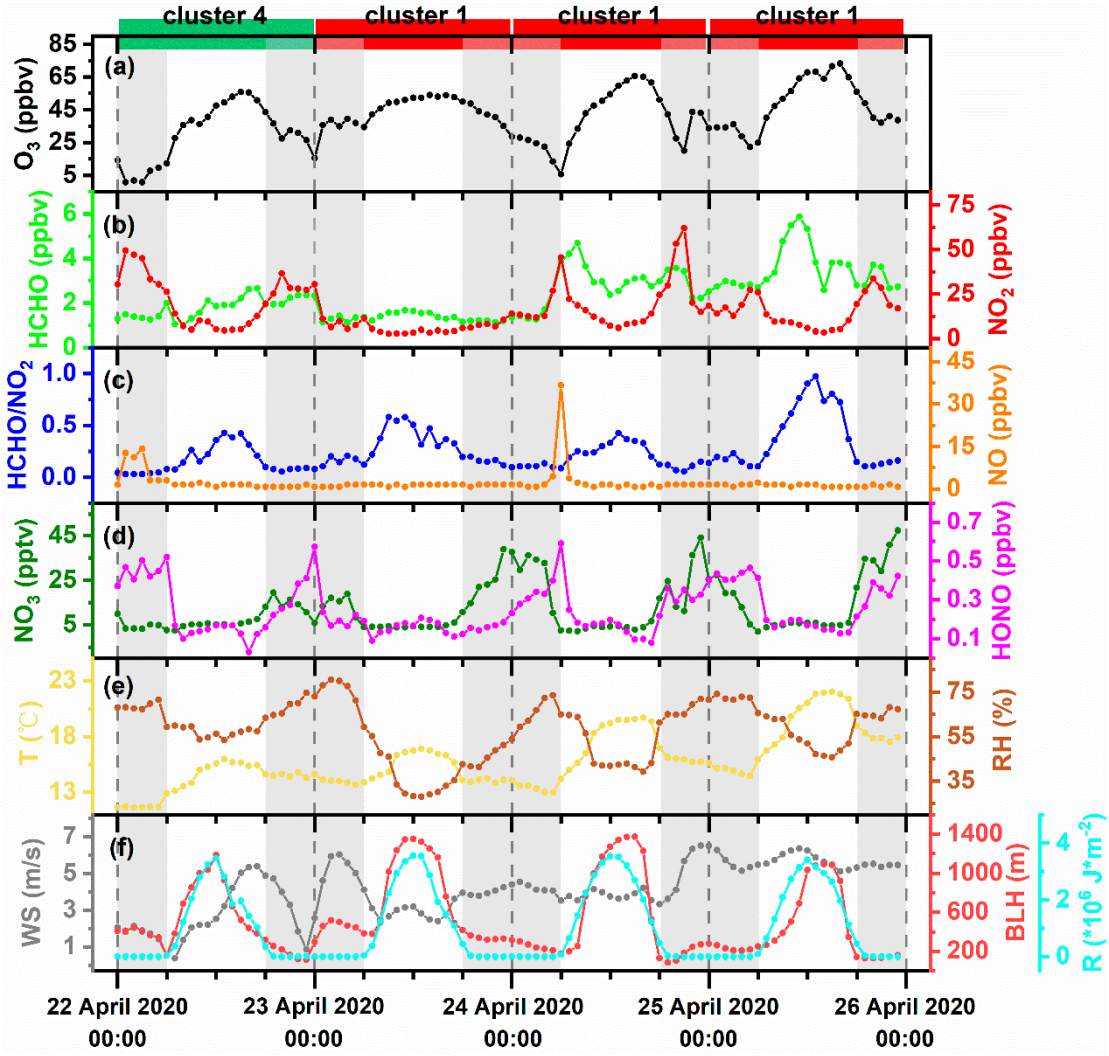

Figure S7. Time series of (a) O_3, (b) HCHO and NO_2, (c) HCHO/NO_2 and NO, (d) NO_3 and HONO, (e) temperature (T) and relative humidity (RH), (f) wind speed (WS), boundary layer height (BLH) and radiation (R).

Table S2. A summary of the relevant atmospheric components for each day.

	Species	April 22	April 23	April 24	April 25
	Cluster	4	1	1	1
After midnight	Averaged O_3 (ppbv)	5.78	33.54	23.95	31.54
	Averaged NO (ppbv)	7.59	1.00	1.74	1.37
	Averaged HONO (ppbv)	0.43	0.26	0.31	0.42
Daytime	O_3 increment (ppbv)	55.19	19.49	60.07	51.01
	Averaged HCHO (ppbv)	1.84	1.45	3.25	4.02
	Averaged NO_2 (ppbv)	9.39	4.39	14.81	9.07
Before midnight	O_3-18:00 (ppbv)	43.52	49.98	51.18	56.08
	Max-NO_3 (pptv)	19.42	38.81	44.06	47.33

High NO was observed after midnight on April 22, which resulted in lower ambient O_3 levels. During the daytime, the meteorological factors of all the 4 days are found to be favorable for the photochemical reaction. The lower concentrations of precursors (both HCHO and NO_2) are observed on April 23, which cause the lower O_3 increment of just 19.49 ppbv. In addition, the correlation coefficients R^2 of the linear fitting for O_3 concentration with corresponding HCHO and HCHO/NO_2 are 0.11 and 0.57, respectively. Compared with HCHO, O_3 reveals a stronger dependence on HCHO/NO_2. The max-NO_3 before midnight increases with O_3-18:00 as shown in Table S2. The boundary layer height keeps lower level before midnight for all the four days, which results in the accumulations of O_3 at the end of daily cycle. For the four days, the response between the O_3 increment and the averaged HONO is unclear. The radiations of the four days are 1.85, 2.05, 2.23, and 2.10 $*10^6$ $J*m^{-2}$, respectively, which are all belonging to the category of lower radiation referred in Figure 8.

Through the case study for a high O_3 pollution episode, we found that the variabilities of O_3, the relative atmospheric components and meteorological factors are consistent with the clustering characteristics.

References

1. P. Govender and V. Sivakumar, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980-2019), *Atmos. Pollut. Res.*, 2020, **11**, 40-56.
2. Y. A. Wijaya, D. A. Kurniady, E. Setyanto, W. S. Tarihoran, D. Rusmana and R. Rahim, Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities, *Tem J.*, 2021, **10**, 1099-1103.
3. J. C. R. Thomas, M. S. Peñas and M. Mora, 2013.
4. A. A. Vergani and E. Binaghi, 2018.
5. H. He, Z. H. Zhao, W. W. Luo and J. H. Zhang, Community Detection in Aviation Network Based on K-means and Complex Network, *Comput. Syst. Sci. Eng.*, 2021, **39**, 251-264.
6. R. Yuan, An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction, *J. Seismol.*, 2021, **25**, 1005-1020.