Supplement of Atmos. Chem. Phys., 22, 15313–15331, 2022
https://doi.org/10.5194/acp-22-15313-2022-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Cluster-based characterization of multi-dimensional tropospheric ozone variability in coastal regions: an analysis of lidar measurements and model results

Claudia Bernier et al.

Correspondence to: Yuxuan Wang (ywang246@central.uh.edu)

Text S1. Description of input features

The K-means clustering algorithm require two inputs: the input features that will be used to cluster the data and the number of clusters. The input features were chosen explicitly based on the structure of the lidar measurements and the structure of the vertical and diurnal pattern of ozone development. The two altitude subsets were chosen to represent the structure of the vertical atmosphere. Therefore, we chose the altitude subset 0 – 2000 meter to represent the complete evolution of the boundary layer, while the 2000 – 4000 meter subset represents the part of the vertical profile in which other factors such as longer range transport of pollutants would be of greater influence. A more in-depth analysis of the development of the boundary layer was out of the scope of this work. The 4 time subsets were chosen to represent the common diurnal pattern of pollutant behavior. Tropospheric ozone has a common diurnal pattern that is greatly influenced by the presence of sunlight. The first subset of time (F1, F5), represent the early morning before the sunlight has reacted with precursor pollutants to create tropospheric ozone. The second time subset (F2, F6) represents the time of day in which the sun rising, and morning traffic have begun which both have an influence on ozone chemical reactions. The third subset of time (F3, F7) represents the midday time in which the sun is at its' full peak. At this time of the day tropospheric ozone usually peaks and remains at the maximum concentration of the day. The final time subset (F4, F8) represents the evening time in which the sun has/begun to set, and ozone concentrations decrease. With the goal of clustering most efficiently, this is in part to simply the data so that the results of the clusters are not weakened by too many input features. But also, not oversimplify the data so that we lose the details of the lidar data.

Text S2. Description of clustering algorithm and cluster efficacy tests

Multiple tests were run on the initial data input features as well as the clustered results to test the efficacy of the clustering. Testing the cluster tendency of a dataset is important before applying a clustering algorithm to determine the cluster ability and efficacy of the dataset. A dataset that has random structures will not contain meaningful clusters. The Hopkins Statistic (Lawson and Jurs 1990) test the cluster tendency by measuring the probability of a dataset has uniform structures. A result higher than 0.50 signifies good cluster tendency and a value higher than 0.75 signifies a high cluster tendency (at the 90% confidence level) (Han, Kamber, and Pei 2012). The results of the Hopkins Statistic for the dataset in this study was 0.78. To visualize the cluster tendency of our dataset and the conclusions of the Hopkins Statistic, we applied the algorithm of the visual assessment of cluster tendency (VAT) approach (Bezdek and Hathaway, 2002) which uses the Euclidean distance measure to compute the dissimilarity matrix in the dataset and creates an ordered dissimilarity matrix image. Figure S2 shows the VAT approach results which indicates high similarity (red) and low similarity (blue) and confirms a cluster structure (not random) within our dataset.

The elbow method, which looks at the percentage of variance based on the number of clusters, was one method used to determine the optimal number of clusters for the dataset (Figure S3). The number of clusters is determined once the percentage of variance drops therefore the addition of clusters will not add much more information. In this work, this happens once at 2 clusters and again 6 clusters. Another approach is the NbClust package (Charrad et al., 2014) in R which offers 30 different indices (e.g., Silhouette, Dindex, Hubert Statistic, etc.)

for choosing the optimal number of clusters. The results of the indices concluded that 8 of the indices proposed 5 clusters as the best number while 4 indices proposed 6 clusters as the best number. Based on the different tests as well as testing the quality of the clustering results using the silhouette method (Kaufman & Rousseeuw, 1990).

The K-means algorithm begins by selecting the objects randomly from the dataset that will serve as the initial cluster centers (centroids or cluster means). After, each other datapoint is assigned to one of the centroids where it is closest to using the Euclidean distance between the centroid and the datapoint. The algorithm computes a new cluster mean for each cluster and using the recalculation, reassesses the assignments reassigning the clusters if needed. These steps are continuously repeated until the cluster assignments remain unchanged, and the clusters are finalized. The results of the K-means clustering algorithm (Figure S4), illustrates the distribution of the data and how the algorithm works in assigning the 6 clusters. The clustering can be further elucidated using principal component analysis (PCA) which demonstrates two of the dimensions in comparison (Figure S5). The PCA function does not show all the dimensions (input features) but rather two principal components that delineate the majority of the variance.

Table S1. Mean normalized bias & Correlation coefficient (R)

a) GEOS-Chem Low-level	Cluster 1 - HMO	Cluster 2 - LLO	Cluster 3 - MCO	Cluster 4 - HLO	Cluster 5 - LMO
Bias	- 0.10	0.07	0.13	- 0.04	- 0.09
R	0.53	0.55	0.51	0.61	0.55
b) GEOS-Chem Mid-level					
Bias	- 0.44	- 0.44	- 0.27	- 0.30	- 0.18
R	- 0.002	- 0.033	- 0.26	0.11	0.23
c) GEOS-CF Low-level					
Bias	0.30	0.50	0.67	0.41	0.45
R	0.74	0.60	0.56	0.61	0.54
d) GEOS-CF Mid-level					
Bias	- 0.22	- 0.07	0.05	0.02	0.28
R	0.51	0.14	- 0.24	0.21	0.74

Table S1. Calculated mean normalized bias and correlation coefficient (R) by cluster. a) Low-level GEOS-Chem, b) Mid-level GEOS-Chem; c) Low-level GEOS-CF and d) Mid-level GEOS-CF results.

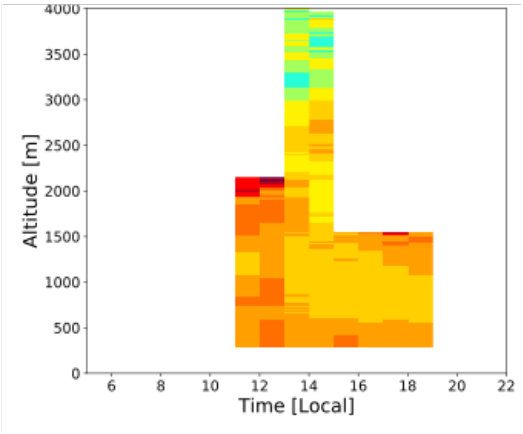

Figure S1. Results from the clustering: Cluster 6 which was assigned only one date (2018-06-17). Considered an outlier and was removed from the analysis.

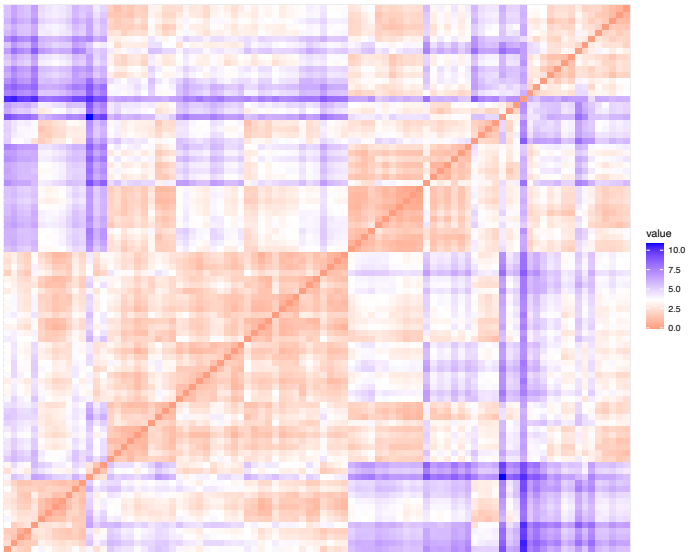

Figure S2. Visual assessment of cluster tendency (VAT) approach. Dataset high similarity (red) and low similarity (blue).

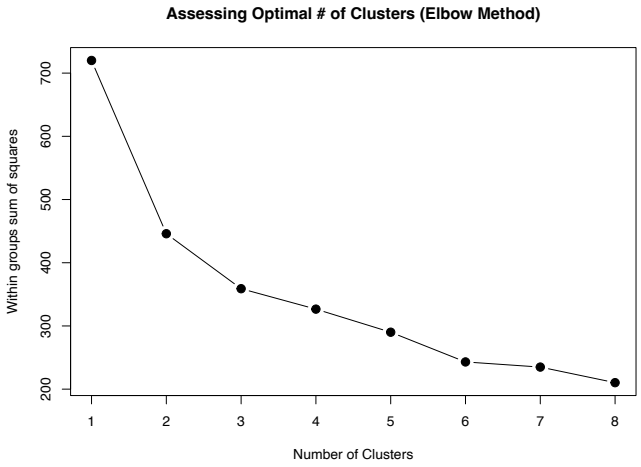

Figure S3. Elbow Method: Assessing optimal number of clusters for dataset.

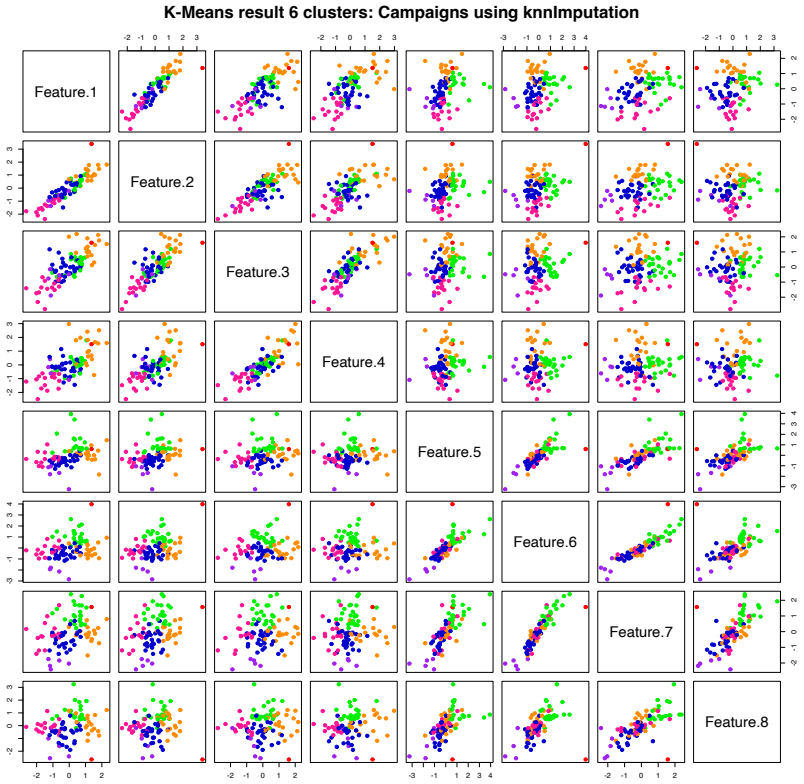

Figure S4. K-means cluster results with original 6 clusters.

Figure S5. Principal component analysis results: two principal components.

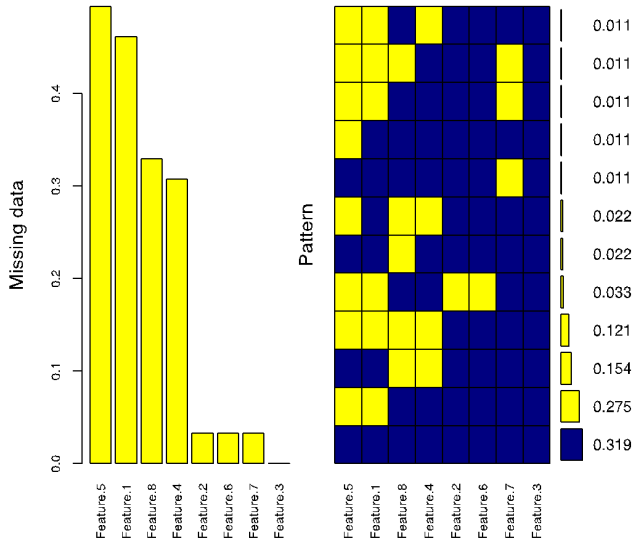

Figure S6. Percentage and pattern of missing data points by each feature used for clustering.

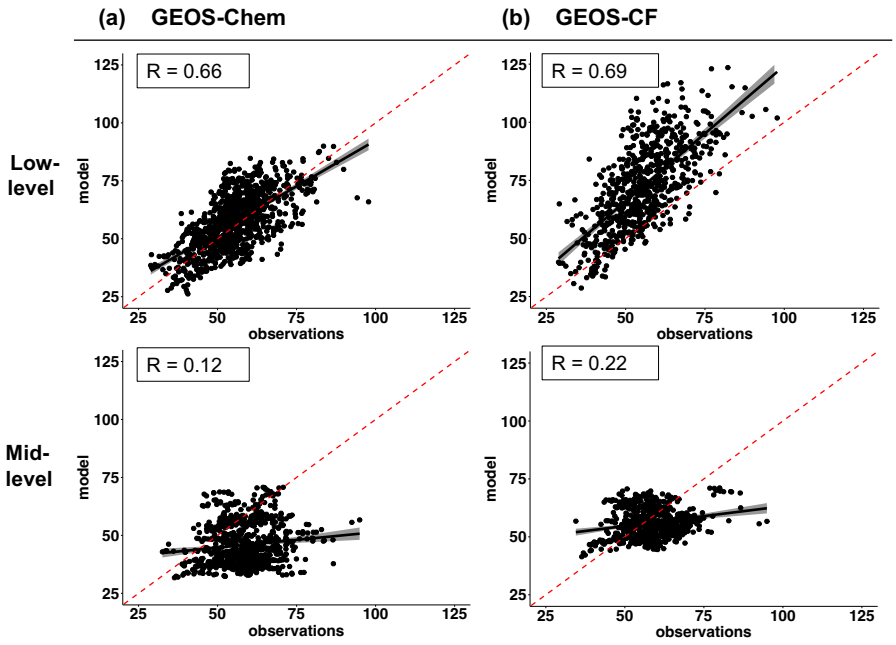

Figure S7. Overall O_3 correlation between the lidar observations versus a) GEOS-Chem and b) GEOS-CF split by low-level (top panel) and mid-level (bottom panel).

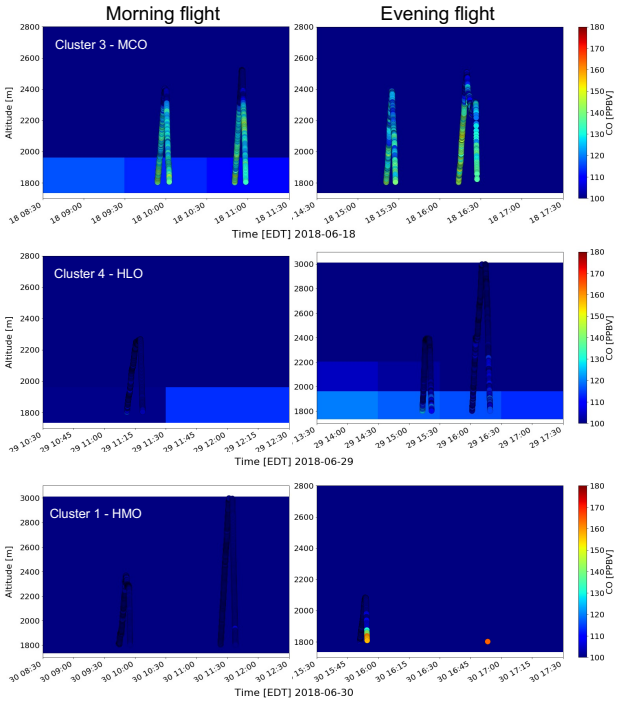

Figure S8. Aircraft measurements and GEOS-Chem simulated CO in the free troposphere during OWLETS-2. Measurements from the UMD Cessna 402B Research Aircraft.

References:

Bezdek, J. C. and Hathaway, R. J.: VAT: A Tool for Visual Assessment of (Cluster) Tendency, Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, 12-17 May 2002, pp. 2225-2230.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A.: NBCLUST: An R package for Determining the Relevant Number of Clusters in a Data Set, Journal of Statistical Software, 61,(6), https://doi.org/10.18637/jss.v061.i06, 2014.

Han, J., Kamber, M., and Pei, J.: Data Mining: Concepts and Techniques. 3rd ed. Boston: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00016-2, 2012.

Kaufman, L. and Rousseeuw, P.: Finding groups in data: An introduction to cluster analysis, New York, Wiley, 1990.

Lawson, R. G., and Jurs, P. C.: New Index for Clustering Tendency and Its Application to Chemical Problems, Journal of Chemical Information and Computer Sciences, 30 (1): 36–41. http://pubs.acs.org/doi/abs/10.1021/ci00065a010, 1990.