Cluster Analysis of Look Rock High Ozone Events and Related Air Parcel Trajectories for 2009-2011

Prepared by Stephen F. Mueller, CCM
Tennessee Valley Authority
September 2012

Cluster analysis is a statistical procedure designed to explore relationships within groups of data. As such its primary use is to identify commonalities across different sets of variables with no assumed prior knowledge of how such variables might be related (Wilks, 2006). Cluster analysis has been applied in studies of atmospheric and climate data. The SAS® statistical software package provides tools for performing cluster analysis. An analysis of Look Rock ozone and trajectory data during 2009-2011 high ozone events was performed using the SAS ACECLUS and CLUSTER procedures[1] (SAS version 9.2).

The ACECLUS procedure analyzes raw data to convert it into variables better suited for cluster analysis. This analysis converts data that are elliptical in *n*-dimensional space (i.e., exhibit more variation in one direction than the other) into spherical data that are then re-configured into *n* canonical (transformed) variables representing different combinations of the original data. In this way, ACECLUS performs a procedure similar to what is done in computing principal components for principal component analysis. The resulting canonical variables are then made available for analysis using CLUSTER.

CLUSTER performs a hierarchical analysis starting with all data points in their own separate cluster. In a continuous sequence, the data points are combined ("agglomerated") into fewer clusters by associating data pairs having the most similarities based on the relative separation between the cluster-mean values of the canonical variables. The relative degree of reduction in total parametric variance across the remaining clusters provides a measure of how the clustering process reduces variations from more to fewer data clusters. SAS CLUSTER documentation provides guidance, based on several statistical metrics, on how to identify significant numbers of clusters. This guidance was used in developing the clustering results described here.

Data

Simulated air parcel coordinates (latitude and longitude) for trajectories arriving at Look Rock during high ozone episodes form the basis of this cluster analysis along with ozone levels measured at Look Rock. HYSPLIT model trajectories arriving at 10, 100 and 500 m above Look Rock were provided for different hours during 23 high 8-hr ozone periods of 2009-2011. Data for 63 sets of trajectories were combined with the maximum hourly and 8-hr average ozone (centered on the maximum hourly ozone level) mixing ratios measured on a given day. A subset of 9 high ozone events--represented by 40 sets of trajectories--were accompanied by other air

[1] Documentation found at
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/ viewer.htm#aceclus_toc.htm (ACECLUS) and
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#cluster_toc.htm (CLUSTER).

quality data to provide a more detailed picture of the conditions associated with the high ozone events. Thus, 14 events were characterized only by ozone and 23 sets of trajectory data.

For the 9 "detailed" events, data provided 8-hr averages and hourly maximum mixing ratios of CO, NO, NO_2, NO_y and photochemical age (τ=NO_z/NO_y), along with 8-hr average solar radiation flux and ridge top wind direction and speed. These 8-hr averages coincided with the 8-hr ozone averages previously mentioned.

Analysis

Trajectory coordinates and ozone mixing ratio (hourly maxima or 8-hr averages) provide excellent variables for cluster analysis because they are linear variables with little covariance between data pairs. The canonical variable equivalents input into CLUSTER were selected for their expected ability to characterize Look Rock high ozone events. Although any and all trajectory coordinates could have been used, it is best to minimize the number of parameters in a cluster analysis for ease of interpretation. To this end, CLUSTER was run with trajectory coordinates for 10 and 100 m trajectories (lat_{10}, lon_{10}, lat_{100} and lon_{100}) at 12 hr upwind of Look Rock in combination with either the 8-hr average (C_{8h}) or maximum hourly (C_{max}) ozone mixing ratio for each event. Both 3 and 5 variable clustering were examined (3 parameter clustering used one ozone metric and one set of trajectory coordinates; 5 parameter clustering used one ozone metric and both sets of trajectory coordinates). Additional parameters add more canonical variance to each analysis (and more potential clusters) and there can be a trade-off between the useful information derived from the clusters and quantity of results to be interpreted.

Although both C_{8h} and C_{max} were tested, neither was clearly superior in providing more definitive clusters. A bit more variance was contained in the fewest significant clusters using C_{max} in combination with both trajectory points. Also, more clusters were needed to identify a major portion of the overall parametric variances when data from both trajectories were used rather than only one of the two trajectories. However, given the complexity of boundary layer airflow and the potential for emissions from different locations to contribute to ozone formation in an environment characterized by complex topography, it was decided that more useful results could be obtained by jointly analyzing the 10- and 100-m trajectory coordinate fields.

After reviewing results based on C_{max}, lat_{10}, lon_{10}, lat_{100} and lon_{100} (a 5-parameter model) it was evident that the 9 event days accompanied by detailed Look Rock data could be grouped into only 5 clusters representing 79 percent of the total parametric variance in the data. However, to achieve a similar degree of parametric variance (80 percent) required 11 clusters for the larger 23 event/63 trajectory data set. Clusters are labeled using C_1, C_2, ...C_n for an *n*-cluster result. The first cluster represents the largest set of agglomerated variance with each subsequent cluster representing decreasing amounts of variance. Comparing trajectories common to both sets of results (i.e., for the 9 event subset and the 23 events) it is possible to identify how well the 9-event subset represents the larger 3-year data base. Table 1 summarizes the commonalities and differences between the two clustering analyses.

There is no perfect one-to-one match between the two sets of clusters. The larger number of clusters needed to characterize the 23-event data set introduces more clusters than could be

identified using the smaller event subset. This causes some trajectories clustered together in the smaller subset analysis to split across multiple clusters in the larger set. However, some clusters remained intact between the two analyses. The most important (largest contribution to total variance) cluster for the 9-event subset, C_1, was contained entirely in the second cluster of the full data set (denoted FC_2). The second subset cluster (C_2) was mostly included in FC_3 although small portions were also found in FC_7 and FC_{10}. All of C_3 mapped to FC_4, all of C_4 mapped to FC_7 and all of C_5 mapped to FC_8. C_6 and FC_{11} also had a large degree of overlap and both included the largest number of individual data points with the smallest relative contributions to total variance within their data sets. Other subset clusters were split across the clusters of the full data set in varying amounts. The complete data set identified 3 clusters--FC_1, FC_5 and FC_6--that had no cross-mapping with the subset. These 3 unmapped clusters represent 37 percent of the total variance in the larger set of data.

Table 1. Relative mapping between clusters derived from the full 3-year 23-event data set and the 9-event subset.

Clusters[a]	
23-Event Set	*9-Event Subset Mapping*[b]
FC_1	
FC_2	C_1 (100%)
FC_3	C_2 (67%)
FC_4	C_3 (100%)
FC_5	
FC_6	
FC_7	C_2 (17%) & C_4 (100%)
FC_8	C_5 (100%) & C_6 (5%)
FC_9	C_6 (14%)
FC_{10}	C_2 (17%) & C_6 (9%)
FC_{11}	C_6 (73%)

[a]The cluster designations in each column refer to the designations used for the individual cluster analyses and are not equivalent across columns.

[b]Values in () denote the fraction of observations in the subset cluster that contributed to the clusters from the larger data set.

To better understand how well the subset represents the full set it is imperative to examine the relative spatial representations of the different clusters regarding air pollutant transport and the comparative ozone values between the two sets of clusters. Plots of 12-hr upwind 10- and 100-m trajectory locations averaged by cluster are illustrated in Figure 1. Except for FC_8, the cluster-average 10- and 100-m trajectory locations at 12 hr upwind from Look Rock fall within 40 km and one direction sector of each other. Therefore, the trajectories at the two levels are mutually consistent and do not indicate airflow from vastly different origins.

Figures 2 and 3 compare average C_{8h} and C_{max} for the data that fall into the clusters of the full data set. Subset clusters that mapped to the larger cluster set do a good job of representing the average cluster C_{8h} (Figure 2) even though this parameter was not included in the clustering procedure. For C_{max}, the match between the two data sets (Figure 3) was excellent as would be expected because C_{max} was used to define the clusters. Other parameters that match well between the two data sets are the upwind distances (relative to Look Rock) of the 10- and 100-m trajectories (Figures 4 and 5).

Interpretation

With 8 of 11 clusters containing detailed data from the 9-event subset it is possible to determine how most subsets differ from the others. Figures 6-8 illustrate cluster averages of 8-hr NO_y, CO

Table 2. Summary of event cluster characteristics.

Cluster	% of Trajec-tories	O_3[a]	NO_y/Age	CO	Transport[b]
1	10	Highest	N/A	N/A	From NNE within 40-45 km (i.e., 10 km ENE of downtown Knoxville)
2	14	High	Moderate/ high	Low	From NNE/NE within 75-80 km
3	6	Moderately high	Moderate/ high	Moderate	From S/SSW within 150-200 km (passing east of Atlanta)
4	3	Moderately high	Moderate/ high	Moderate	From S within 185-200 km
5	3	Moderate	N/A	N/A	From NE within 100-150 km
6	11	Moderately low	N/A	N/A	From NNE/NE within 70-80 km
7	5	Moderately low	Moderately high/ high	Moderate	From SE within 145-160 km
8	8	Moderate	High/low	Moderate	From ESE through SSW sectors within 15-35 km
9	5	Moderately low	Low/high	High	From WSW within 135-140 km (Chattanooga)
10	5	Moderately low	Moderate/ moderate	Moderate	From SW within 80-100 km
11	30	Moderately low	Moderate/ moderate	Moderate	From WSW within 75-80 km

[a]Ratings are relative across the population of all event clusters. All 8-hr average ozone levels exceed 75 ppb except for cluster 6 (74 ppb).
[b]Distances refer to transport during the 12 hours preceding arrival at Look Rock and include information from both the 10- and 100-m trajectories.

and photochemical age (τ). Of those clusters containing data, FC_8 had the highest NO_y levels while FC_9 had the lowest. The airmasses in FC_9 had the highest mean τ while the lowest was found for FC_8. In general, τ was inversely associated with NO_y. FC_9 had the highest average CO levels and FC_2 had the lowest. From these data the clusters can be characterized as summarized in Table 2.

Clusters 1, 2 and 6

The highest ozone was associated with clusters FC_1 and FC_2 and the transport directions were very similar with 12-hr origins being farther upwind for FC_2 (i.e., winds were stronger for FC_2).

Detailed air quality data other than ozone were not available for FC_1, but given the similarities in transport directions and ozone levels (with lower ozone levels associated with higher transport speeds as expected) it is likely that the air quality signatures of these two clusters are similar. FC_2 had moderate NO_y levels, high τ and low CO. Thus, the airmasses associated with FC_2 do not have a strong urban signature despite passing only 10 km or so east of downtown Knoxville. Instead, they appear more strongly associated with transport from more distant sources despite the fact that 12-hr transport distances were not that high. FC_6 transport overlapped completely with that of FC_1 and FC_2 with speeds intermediate between FC_1 and FC_2. FC_6 ozone levels were lower than either FC_1 or FC_2 and, like FC_1, FC_6 did not have detailed air quality data. Thus, all that can be deduced is that conditions (either meteorology or ozone precursors) were not as amenable to ozone formation in cluster 6.

Cluster 5

Cluster 5 is set apart from the other clusters by the strength of its transport and the direction (NE) of its origins. With air originating in northeast Tennessee during the 12 hours preceding arrival at Look Rock, it is most likely that FC_5 ozone levels were influenced by sources near Kingsport, TN. No detailed air quality data were available.

Clusters 3 and 4

Clusters 3 and 4 are unique in that they have the highest transport speeds during the 12 hours preceding arrival at Look Rock and originate from well south of Look Rock over north central Georgia. Ozone levels are relatively high, NO_y and CO levels are moderate, and τ is high. These clusters could be associated with emissions east of Atlanta and are definitely not due to emissions anywhere near Look Rock.

Clusters 7 and 8

Clusters FC_7 and FC_8 share several features--both originate southeast of Look Rock, both have moderately high ozone and CO levels and both have relatively high NO_y levels. The primary differences are in transport speeds/distances and τ. FC_7 is associated with very aged air (high τ) and high transport speeds; 12-hr airmass origins are over extreme northeast Georgia and extreme western South Carolina. However, FC_8 is associated with the lowest transport speeds of any cluster and the largest directional divergence between 10- and 100-m trajectories. Airmass origins are within 30 km of Look Rock, essentially over the Great Smoky Mountains National Park (GSMNP) and nearby national forests and wilderness areas. There are no significant source regions in this area. Despite this, τ was lower for FC_8 than any other cluster and this implies a nearby source of NO. Given the transport directions and low winds it is likely that Look Rock air quality in cluster 8 was influenced by emissions from the small Happy Valley community east of the Look Rock/Foothills Parkway (Chilhowee Mountain) and west of GSMNP and, perhaps to some extent, by the Look Rock campground east of the air monitoring station. Emissions from fireplaces, wood stoves and campfires could contribute the relatively fresh NO emissions and moderate CO levels that were measured.

Clusters 9 and 11

Clusters FC_9 and FC_{11} share a common transport direction (WSW) and moderately low high ozone levels. However, they differ considerably in their transport speeds and other air quality characteristics. FC_9 has low NO_y but high τ and CO. Given the 12-hr transport distances and directions it is likely that FC_9 events were influenced by Chattanooga emissions. The relatively high CO levels may reflect the fact that FC_9 trajectories follow the Chattanooga-Knoxville transportation and industrial corridor for much of the 12-hr period prior to arrival at Look Rock. FC_{11} events follow similar paths but transport speeds are lower and both CO and τ are lower than for FC_9. The lower FC_{11} speeds put the Chattanooga source region further upwind in time than for FC_9. The lower τ for these events implies that more nearby sources are important than for FC_9 events and the lower CO suggests that they are not mobile sources. FC_{11} comprises the largest single cluster with 30 percent of all computed trajectories.

Cluster 10

Trajectories for cluster 10 originate in northwest Georgia 12 hours upwind of Look Rock. This region does not have large urban areas but it does include a variety of industrial facilities and electric generating stations. Air quality data suggest that these moderately low high ozone events are associated with moderate levels of emissions.

Summary

A cluster analysis of high ozone events at Look Rock provides insight into the sources and conditions associated with various categories or clusters of events. The data suggest that, out of 11 distinct clusters, only 4 (FC_1, FC_2, FC_5 and FC_6) are likely to have any contributions from the Knoxville urban area. Of these 4, one (FC_5) has such strong transport from the northeast that airmasses spend very little time over Knoxville-area emissions sources. Of the remaining 3 clusters, FC_1 has the highest ozone levels and the greatest potential for Knoxville-area contributions because of the relatively long time spent by sampled air parcels traveling across the urban source region. Clusters 1, 2 and 6 with the highest potential impacts from the Knoxville area comprise 35 percent of the computed air parcel trajectories associated with Look Rock high ozone events. In other words, 65 percent of the Look Rock high ozone events are not associated with or significantly impacted by Knoxville-area emissions.

Reference

Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*, Academic Press, New York, 2006, 549-559.

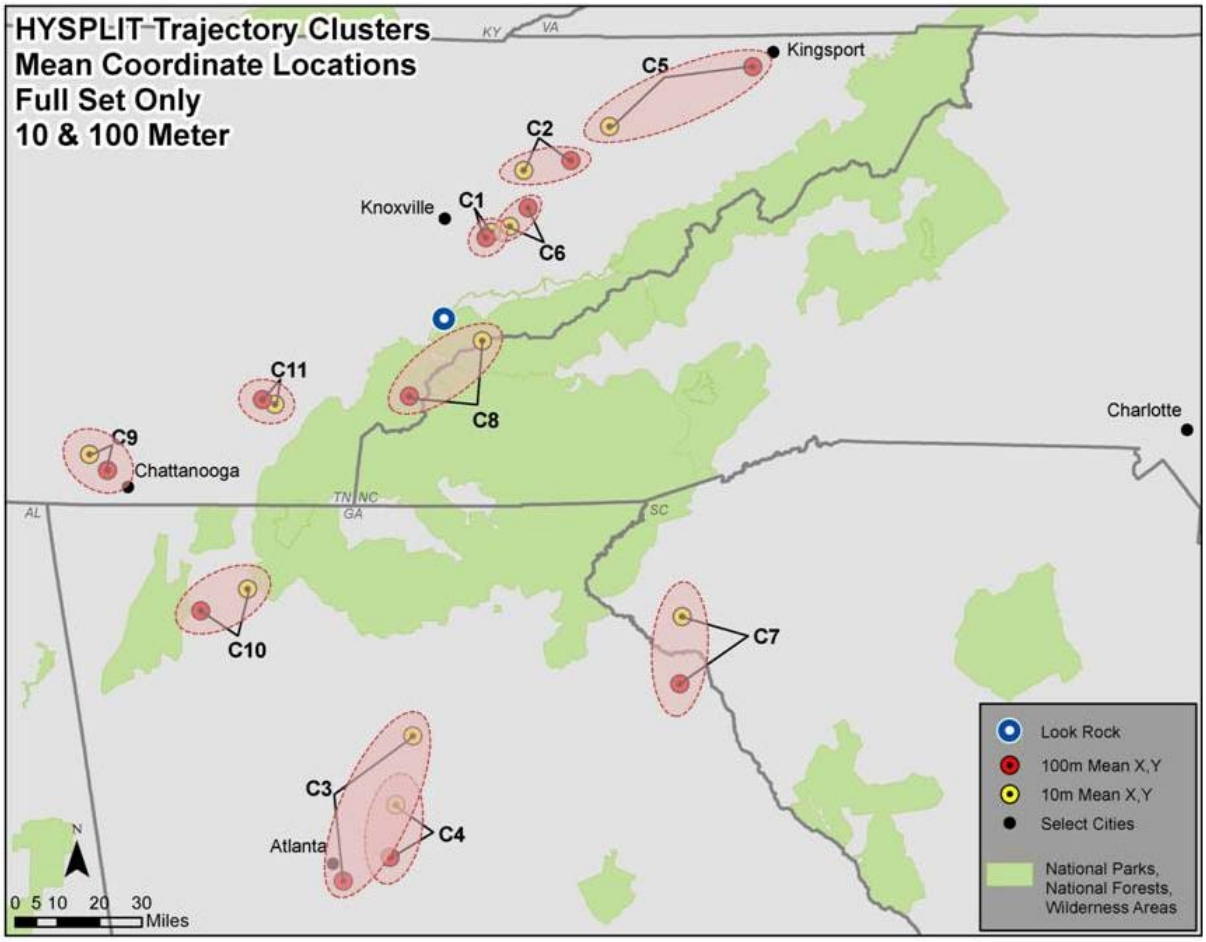

Figure 1. Cluster-averaged 10- and 100-m trajectory locations at 12 hours upwind of Look Rock during high ozone events. Locations are labeled by cluster number. Pink ovals denote ozone precursor source regions at 12 hours prior to arrival at Look Rock.

Figure 2. Comparison of cluster averaged 8-hr ozone mixing ratios for periods centered on the maximum observed one hour ozone value on Look Rock high ozone event days. Note that this is not the same as the daily maximum 8-hr average ozone value.

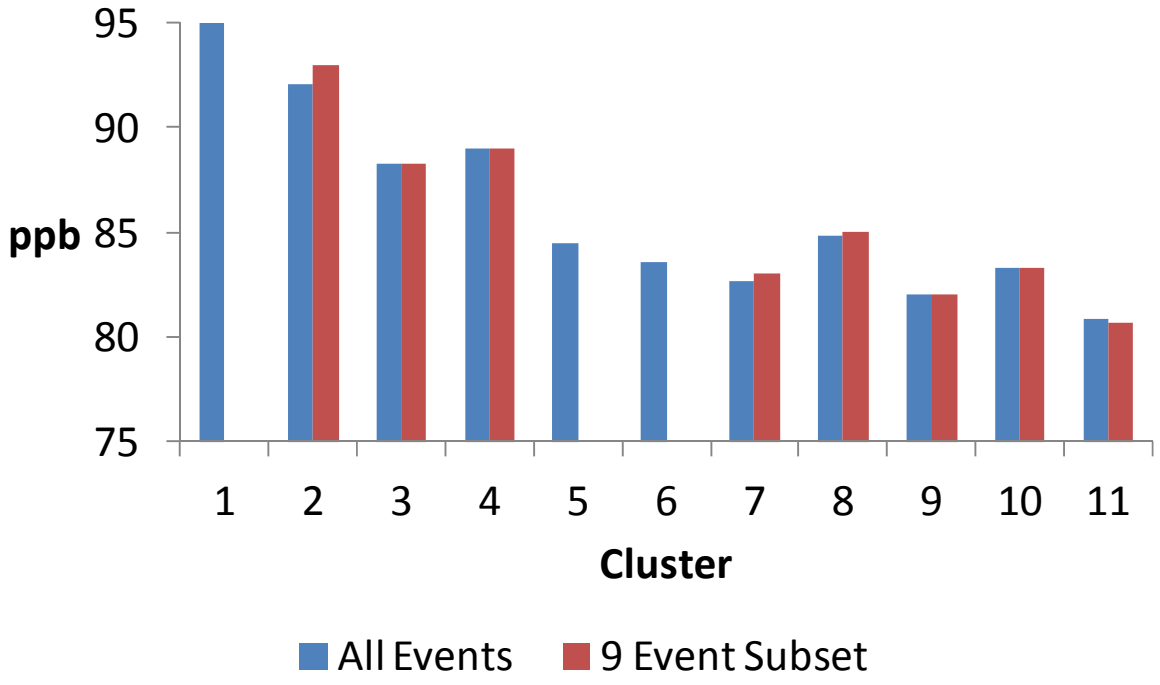

Figure 3. Comparison of cluster averaged maximum hourly ozone mixing ratios observed at Look Rock on high ozone event days.

Figure 4. Cluster average computed trajectory distances upwind of Look Rock at 12 hr preceding arrival at Look Rock during high ozone events for trajectories arriving at a height 10-m above the destination.

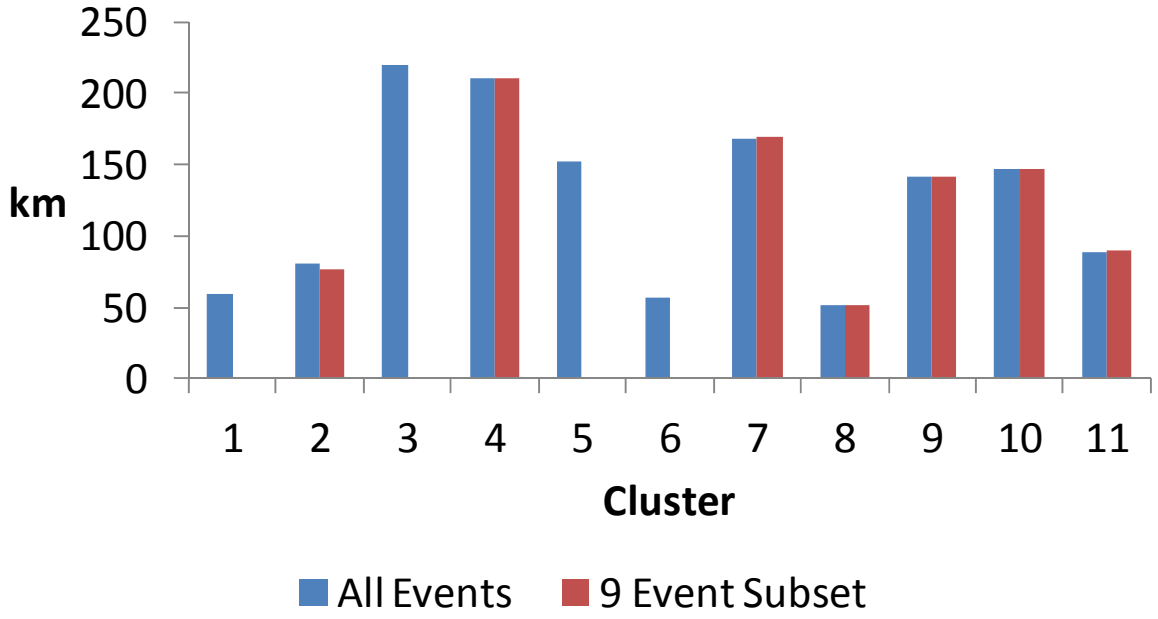

Figure 5. Cluster average computed trajectory distances upwind of Look Rock at 12 hr preceding arrival at Look Rock during high ozone events for trajectories arriving at a height 100-m above the destination.

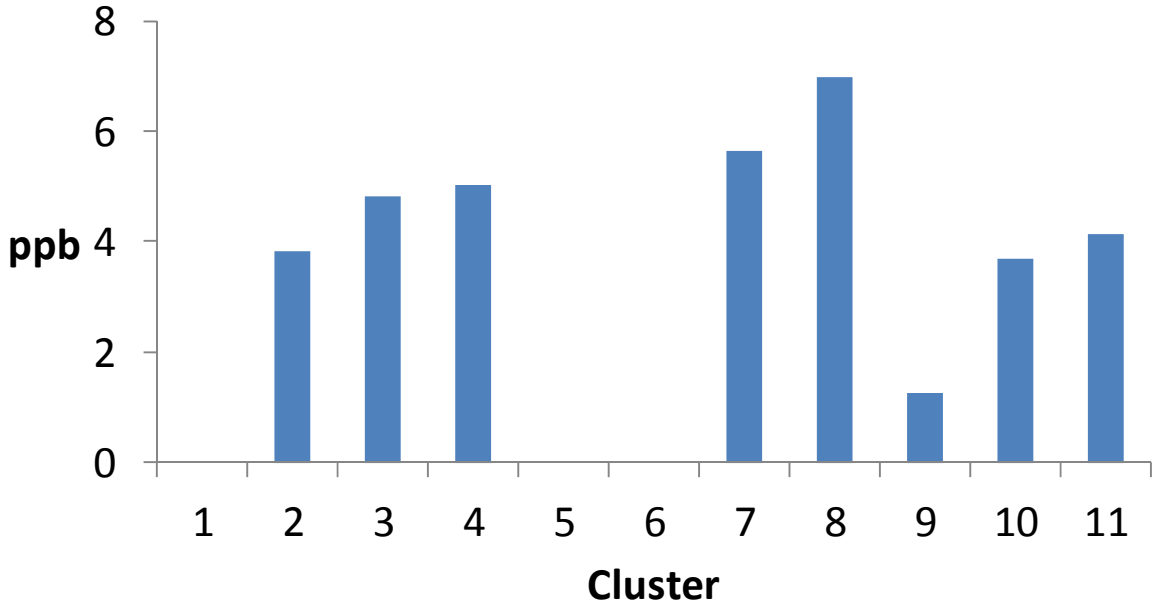

Figure 6. Cluster average NO_y mixing ratios for 8-hr periods centered on the daily maximum hourly ozone mixing ratio measured during high ozone events at Look Rock.

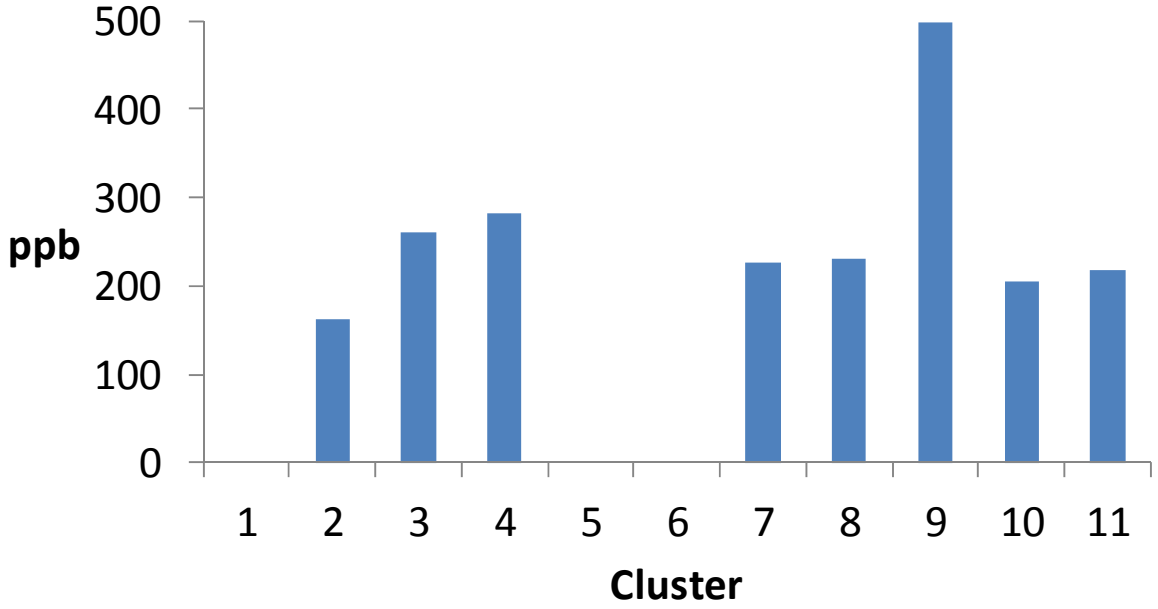

Figure 7. Cluster average CO mixing ratios for 8-hr periods centered on the daily maximum hourly ozone mixing ratio measured during high ozone events at Look Rock.

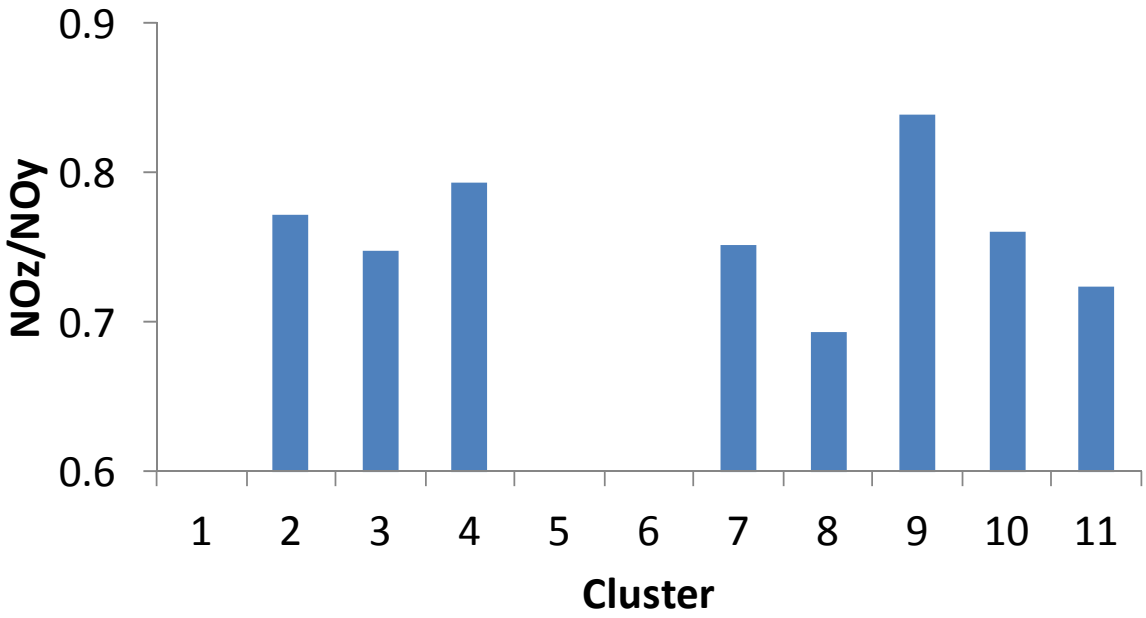

Figure 8. Cluster average photochemical age (τ) for 8-hr periods centered on the daily maximum hourly ozone mixing ratio measured during high ozone events at Look Rock.