Meteorology-induced variations in the spatial behavior of summer ozone pollution in Central California

Ling Jin[1],
Robert A Harley[1, 2]
Nancy J Brown[1]
1. Atmospheric Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
2. Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720

Corresponding author: Nancy J Brown, njbrown@lbl.gov, 510-486-4241, fax 510-486-5928

Keywords: Summer Ozone, Central California, Cluster Analysis, Photochemical Modeling

Abstract

Cluster analysis was applied to daily 8 h ozone maxima modeled for a summer season to characterize meteorology-induced variations in the spatial distribution of ozone. Principal component analysis is employed to form a reduced dimension set to describe and interpret ozone spatial patterns. The first three principal components (PCs) capture ~85% of total variance, with PC1 describing a general spatial trend, and PC2 and PC3 each describing a spatial contrast. Six clusters were identified for California's San Joaquin Valley (SJV) with two low, three moderate, and one high-ozone cluster. The moderate ozone clusters are distinguished by elevated ozone levels in different parts of the valley: northern, western, and eastern, respectively. The SJV ozone clusters have stronger coupling with the San Francisco Bay area (SFB) than with the Sacramento Valley (SV). Variations in ozone spatial distributions induced by anthropogenic emission changes are small relative to the overall variations in ozone amomalies observed for the whole summer. Ozone regimes identified here are mostly determined by the direct and indirect meteorological effects. Existing measurement sites are sufficiently representative to capture ozone spatial patterns in the SFB and SV, but the western side of the SJV is under-sampled.

1 Introduction

Ozone is a designated criteria pollutant and has harmful effects on human health (Bell and Dominici 2008) and agriculture productivity (Emberson *et al.* 2009). Understanding the spatial distribution of ozone pollution and its temporal changes is of interest to determine the level of exposure. Meteorological factors can either alter directly ozone chemistry and transport, or indirectly affect ozone formation by influencing light and temperature sensitive precursor emissions from biogenic sources (Dawson et al. 2007; Tao et al. 2007). Thus temporal variations in the magnitude and spatial extend of ambient ozone pollution may reflect the changes in meteorological conditions. Past observational studies identified reoccurring spatial distributions of ozone to determine the associated meteorology that influences air quality (e.g. Ludwig *et al.* 1995; Dayan and Levy 2002; Beaver and Palazoglu 2006). This approach is valid only when changes in ozone are not significantly influenced by factors other than meteorology (e.g. large wild fire emissions and changes in anthropogenic emissions). Furthermore, questioning whether the limited spatial coverage of measurement locations can adequately capture regional level ozone patterns has largely been ignored.

Cluster analysis (Everitt, 1993) is commonly employed to classify pollution regimes, and is traditionally applied to observational data that generally have extensive temporal coverage to account for uncertainties resulting from inter-annual variability and emission changes. When the approach is applied to an extensive modeling study, some of the limitations associated with observational studies can be controlled. Factors other than meteorology such as emission inputs can be specified and contributions from changes in meteorology vs anthropogenic emissions can be compared with sensitivity studies. There are no spatial gaps in the gridded model data, hence spatial representativeness is no longer an issue. Nevertheless, few analyses of photochemical modeling results have been guided by this approach. This is because model development and input data preparation generally limit highly resolved simulations to short modeling periods of a few days which are insufficient for resolving representative patterns.

California's San Joaquin Valley suffers from serious ozone air pollution problems due to its unique geography as well as diverse emission sources from both local and upwind areas. As previous modeling studies have often focused on ozone episodes that last for 3 to 5 days,

knowledge of meteorological regimes that influence ozone behavior has been obtained mainly from statistical analysis of historical observations at limited measurement sites (e.g. Ludwig *et al.* 1995, Beaver and Palazoglu 2009). Photochemical modeling has been conducted for all of summer 2000 in Central California (Jin et al., 2010), which provides an opportunity to demonstrate cluster analysis applied to model results.

In this paper, we characterize regimes of modeled ozone spatial behavior in the San Joaquin Valley in relationship with upwind air basins in Central California, i.e. Sacramento Valley and San Francisco Bay area. Variations induced by changes in anthropogenic emissions are quantified to determine whether ozone regimes identified here are mainly attributable to them or to underlying meteorological processes. Spatial representativeness of existing measurement locations is also discussed. The results can serve as a basis to investigate associated meteorological features to reveal mechanisms for regional ozone transport and accumulation in subsequent work

2 Methods

Cluster analysis and principal component analysis are employed in this study to determine and interpret spatial distributions of ozone in Central California.

2.1 Principal Component Analysis

Both ozone and meteorological fields have large spatial dimensions, for example, 10^4 dimensions are needed to describe surface 8 h ozone maxima on one day for a domain of size 100 by 100. These dimensions are redundant and can be reduced by considering spatial correlation. To accomplish this, Principal Component Analysis (PCA) (Pearson 1902; Hotelling 1935) or Empirical Orthogonal Function (EOF) technique (Lorenz 1935), has been frequently employed by linearly combining the original large number of dimensions into a few new dimensions without compromising much of the data variance. Reduced dimensionality makes the dataset much easier to manage and interpret. The basic idea is that the spatial data at any given time can be represented by the sum of the mean pattern and multiples of a few principal

spatial patterns (PCs or EOFs). These principal component multipliers (defined as "scores") can be used as a reduced set of descriptors for the spatial patterns. The resulting PCs are associated with distinct physical processes when they strongly influence the spatial variations. In this case the scores can be interpreted as indicating the strength of these physical processes.

The spatial data are the 8 h ozone maxima for a specific region, and here these are used to form the data matrix (M) as follows:

$$M = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1P} \\ X_{21} & X_{22} & \cdots & X_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{NP} \end{pmatrix} \quad (1)$$

where, X_{is} is the maximum 8 h ozone concentration on day i at grid cell s; N is the total number of days, P is the total number of grid cells. The anomaly matrix (M') is formed as

$$M' = \begin{pmatrix} X'_{11} & X'_{12} & \cdots & X'_{1P} \\ X'_{21} & X'_{22} & \cdots & X'_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ X'_{N1} & X'_{N2} & \cdots & X'_{NP} \end{pmatrix} \quad (2)$$

where X'_{is} is the departure of X_{is} from the time average at grid cell s. The principal components are determined through singular value decomposition (SVD) on M':

$$M' = UDV^T \quad (3)$$

where U (with dimension $N \times r$, with r the rank of M') and V (with dimension $P \times r$) are orthogonal matrices, whose column vectors consist of left and right singular vectors, respectively. D is a diagonal matrix with singular values d_{ii}. PCs are thus the column vectors in V, and the variance explained by the i^{th} PC is $d_{ii}^2 \Big/ \sum_{j=1}^{r} d_{jj}^2$.

The PCs provide a new orthogonal basis for the daily ozone (8 h) maxima. Ozone maxima on day i: $\vec{X}_i = (X_{i1}, X_{i2}, \cdots, X_{iP})^T$ can now be represented as follows:

$$\begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iP} \end{pmatrix} = \vec{\bar{X}} + \sum_{m=1}^{N_{pc}} a_{im} \begin{pmatrix} x_{m1} \\ x_{m2} \\ \vdots \\ x_{mP} \end{pmatrix} \tag{4}$$

where, $\vec{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N} \vec{X}_i$ is the average ozone maxima over all N days; a_{im} is the score (or coefficient) of the concentration anomaly $\vec{X}_i' = \vec{X}_i - \vec{\bar{X}}$ of day i projected onto the m^{th} PC: $(x_{m1}, x_{m2}, \cdots, x_{mP})^T$. The PCs in Equation (4) are arranged in descending order of importance as defined by the explained variance. Putting all the days together, scores on the first PC have the largest variability, scores on the second PC are next, followed by the third, and so on.

2.2 Cluster Analysis Methods

In this study, cluster analysis is used to define the spatial patterns of daily 8 h ozone maxima of specified subregions in Central California. Days that have similar spatial 8 h maximum ozone, with respect to magnitude and spread, are grouped together to define a unique ozone pattern.

The dissimilarity measurement (D_{ij}) between any two days i and j is the Euclidean distance between the 8-hr maximum ozone vectors ($\vec{X}_i$ and $\vec{X}_j$) associated with them:

$$D_{ij} = \sqrt{\sum_{s=1}^{P} (X_{is} - X_{js})^2} \tag{5}$$

where s is the grid index within the subregion, and P is the total number of grid cells.

In most observational studies, the data matrix is scaled to weigh all the measurement locations equally in the analysis. Observations from each station are scaled independently by removing their mean over all the days then dividing by their standard deviation. In this study, all the grid cells are included in the analysis; hence such scaling is no longer appropriate. Low ozone levels are predicted for the majority of the domain grid cells, while high ozone levels are concentrated near the source regions. Consequently, standardizing each grid cell tends to exaggerate the ozone variations at locations that are relatively clean. No scaling is applied here so that the ozone patterns are distinguished by their absolute ozone magnitude and spatial extent.

In this study, we employ the most widely used partitioning method (the Hierarchical and Partitioning methods are described in Supporting Information), k-means clustering (Anderberg 1973). The clustering procedure partitions the ozone days into k groups, where the within-group sum of square over all the dimensions is minimized. For any given *k*, different initial partitioning gives different final clustering results. In order to stabilize the final clustering result, we repeat the process with 50 random starts and pick the one that has the minimum within-group sum of squares (WSS) (See Supporting information Figure S1) as the final result. This gives us more consistent clustering results.

There is no consensus on an objective method for determining the number of groupings, *k* (Everitt *et al.* 2001). Some studies combined *Hierarchical* methods with *Partitioning* methods to aid visual inspections of the data structure for selecting an appropriate number of groupings (e.g., Kaufmann and Whiteman, 1999; Beaver and Palazoglu, 2006) or used a criterion with single metrics like the Silhouette[1] measure of Kaufman and Rousseeuw (1990). Others were entirely subjective (e.g. Ludwig *et al.*, 1995) so as to have enough clusters to encompass the range patterns in the data set but not too many to reduce greatly the size of each cluster. Although an automated scheme is desirable, some expert knowledge can be incorporated a *priori*, as practiced in Darby (2005), to insure that important information is revealed without engaging in over-interpretation at the same time. Increasing *k* for the k-means method generates more compact clusters and thus generally decreases the within-group sum of squares. To avoid over-interpretation, the difference between WSS_k and WSS_{k+1} is calculated for a number of clusters k = 1, 2, …, 20. WSS generally does not decrease significantly after *k* is in the range 4-7. Therefore, $\{4,5,6,7\}$ forms a candidate set for *k*. The final value of *k* is selected after visually examining

[1] For each observation *i*, the *silhouette number s(i)* is defined as follows: Let *a(i)* = average dissimilarity between *i* and all other points of the cluster to which *i* belongs (if *i* is the *only* observation in its cluster, *s(i) := 0* without further calculations). For all *other* clusters C, put *d(i,C)* = average dissimilarity of *i* to all observations of C. The smallest of these *d(i,C)* is *b(i)* and can be seen as the dissimilarity between *i* and its "neighbor" cluster, i.e., the nearest one to which it does *not* belong. Finally,

$$s(i) = (b(i) - a(i)) / max(a(i), b(i)).$$

Therefore, silhouette number can be used for *describing* how well a data point fits into the cluster that it is assigned to.

cluster centers to insure important spatial patterns are revealed. Silhouette numbers are also computed for the range of *k* that was considered, and they usually do not differ significantly.

2.3 Data and subregions

Hourly ozone concentrations were simulated using the Community Multiscale Air Quality (CMAQ) modeling system for the period from 2 June to 26 September 2000. Details of model inputs and evaluation against observations can be found in Jin et al (2010). The model is capable of reproducing observed ozone concentrations, especially for the San Joaquin Valley.

Clustering of ozone spatial distributions is done for the San Joaquin Valley as well as two major upwind air basins in Central California (see Figure 1), which also experience frequent ozone exceedances during the summertime. Subregions are defined by extending each of three air basins: San Francisco Bay Area (SFB), San Joaquin Valley (SJV), and Sacramento Valley (SV), to include their respective downwind areas (indicated in Figure 1 by lighter colors), according to the dominant summer flows. Including these additional downwind areas allows for capturing of pollutant spatial patterns that are closely related to prevailing wind regimes.

The 8 h ozone maxima were determined for each grid cell and for each day. Days with less than 0.1% of the grid cells in the CCOS domain exceeding 84 ppb, are considered as very clean. These very clean days are labeled as cluster 0, and are excluded from further clustering analysis. The Sacramento Valley is heavily influenced by large forest fire emissions from Aug 18th to 29th. These days are also excluded from clustering analysis because their variations are largely induced by emission changes rather than meteorology, as the meteorological changes induced by the fires were not captured by the model.

3 Results

3.1 Spatial ozone behavior in Central California

PCA was performed for the 8 h ozone maxima for all the days in each subregion to reduce the number of dimensions needed to describe ozone behavior on each day. PCA provides

insights on underlying physical processes; if one or two processes are particularly influential, they are very likely to be reflected in one of the first few PCs. Figure S2 (Supporting Information) reveals that the first six PCs explain more than 90% of the variance in daily ozone maps, and the first three PCs usually account for about 85% of the variance for all subregions. A subspace spanned by the first few PCs can represent the spatial ozone anomalies quite well.

Daily 8 h ozone maxima in each subregion can then be described by the summer average and strengths of individual PCs. Figure 2 presents the 8 h ozone maxima averaged over the whole summer (except for SV, where the days significantly affected by forest fire emissions are excluded). The summer averages (the mean field) describe ozone features common to all the summer days. For example, the SJV has the highest ozone levels among all the air basins, and especially high ozone levels are seen in Fresno, Bakersfield, and Sacramento. Figure 3 shows ozone loadings on the first three PCs for the three subregions. The first three PCs characterize the most prominent patterns observed in the remaining ozone anomaly fields (deviation of daily ozone maxima from the mean field). The anomaly fields describe features unique to individual summer days. Color scales in Figure 3 were chosen to reveal the spatial patterns for individual PCs.

PC1 is mostly positive, describing an average spatial trend in ozone concentrations, governed by the relative distribution of emissions, with higher ozone levels near emission sources and lower in remote rural areas. Higher score on this PC implies higher ozone concentrations with magnitude mostly proportional to the availability of precursors. For example, Fresno ozone increases on hotter days, and this increase is greater than those found for more remote areas. PC2 (and PC3) in each subregion is usually a spatial contrast[2] or gradient in a specific direction, characterized by the transition from positive loadings in one part of the region to negative loadings in another. For example, PC2 for the SJV is a spatial contrast between the northwestern and southeastern sides of the valley, and larger scores on this PC imply larger differences (stronger contrast) in ozone anomalies between the two sides of the valley. These spatial contrasts in ozone concentrations could be caused by meteorological factors such as a temperature gradient under certain synoptic conditions, or wind flows that carry pollutants from source regions to downwind areas. In the case of SJV PC2, for example, the spatial contrast

[2] Contrast is a statistical term usually used in experimental designs. The coefficient (or score, here) of a contrast measures the average difference between treatments that are marked by different signs in the contrast.

follows the direction of typical up-valley flows, moving pollutants from the source regions in the valley to the Sierras and the Mojave desert, which causes a decrease in ozone levels in the northwest and an increase in the southeast. A summary description of the first three PCs for each of the subregions can be found in Table 1

Cluster analysis can be performed with the original ozone data, or with the reduced data set by projecting ozone anomalies onto the subspace defined by the first six PCs that generally capture more than 90% of the variability in the original data set. Both methods produce the same results. Clusters are ranked according to the average scores on the first PC (usually accounting for ~70% of total variance), which indicates the magnitude of the average ozone levels for that cluster.

3.2 Ozone regimes in the SJV

The San Joaquin Valley has the most serious ozone problems in Central California. After excluding 27 very clean days (defined in section 2.3 as cluster 0), 90 days are grouped into six ozone clusters for the SJV. The distribution of cluster members is visualized in a reduced dimension space (Figure 4), while average ozone anomaly fields of each cluster are plotted using the original dimensions (Figure 5). Average ozone anomaly fields obtained by subtracting the seasonal mean provide a better representation of the spatial features that are unique to individual clusters. Figure 4 is a boxplot[3] that summarizes the distribution of the scores on the first three PCs to illustrate how each cluster (1 to 6) is significantly different from the others in terms of one or more PC scores.

Clusters 1 and 2 have significantly less ozone than other clusters, as indicated by negative scores on PC1, and they differ from each other by spatial contrasts mainly described by PC2. The spatial contrasts are also reflected in the mean ozone anomaly maps of these two clusters:

[3] Boxplot is a box and whisker plot for scores on each PC. The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. If there are no data outside the whisker, a dot is placed at the bottom whisker. In a notched box plot the notches represent a robust estimate of the uncertainty about the medians for box-to-box comparison. Boxes whose notches do not overlap indicate that the medians of the two groups differ at the 5% significance level.

elevated ozone on the western side of the valley in cluster 1 and on the eastern side in cluster 2. **Clusters 3 to 5** have moderate ozone levels with similar PC1 scores, higher than clusters 1 and 2, and lower than cluster 6. These three clusters differ from each other based on spatial distributions of ozone anomalies governed by PC2 and PC3. Cluster 3 has higher scores on PC3 compared to clusters 4 and 5. Clusters 4 and 5 differ by having PC2 scores of opposite signs. These features are in accord with spatial patterns of mean ozone anomalies (see Figure 5): Cluster 3 has elevated ozone levels on the north side, Cluster 4 on the western side, and Cluster 5 on the more downwind southeastern side. **Cluster 6** has significantly higher ozone than others (highest PC1 scores).

The dynamic relationships among the SJV clusters can be revealed by modeling the transitional probabilities (e.g. Beaver *et al.* 2008); however their statistical significance cannot be determined due to the number of days (117) considered in this study. The number of days when ozone air quality transitioned from one cluster to another (see Supporting information Table S1) indicates that low-ozone days are less likely to develop into high-ozone clusters on the next day. This implies that ozone accumulation in the SJV usually takes multiple days. The transition matrix indicate the two most probable paths for SJV ozone levels to develop from low to high: 1→4→6 (then back to 5), or 2→3→5 (then back to 2, or, less likely, develop into 6). As seen in Figures 4 and 5, clusters 2 and 5 have similar spatial contrast features as characterized by PC2, but different average ozone levels as characterized by PC1. Similarly, clusters 1 and 4 have similar spatial contrasts but different ozone levels. The two paths (1→4→6 and 2→3→5) may imply the development of two different weather regimes.

3.3 Ozone regimes in the SFB and the SV

The Bay area and the Sacramento Valley are, at times, influential upwind source areas that affect air quality in the San Joaquin Valley. The remaining 90 days from summer 2000 (after clean days are excluded) are grouped into four SFB ozone clusters. Five clusters are obtained for the SV after excluding 12 additional days on which air quality was significantly affected by forest fires.

The average spatial patterns of ozone anomalies indicated in Figures 6 and 7 are in accord with the PC scores of the SFB and SV (Figure S3 in Supporting Information). SFB cluster 4 has

more elevated ozone levels near the ocean (western and southern) rather than in more inland areas as in cluster 3. This is likely due to a decrease in strength of the sea breeze. SV cluster 3 experiences an ozone decrease in Sacramento and an increase in the northern end of the valley. With ozone transported to downwind areas, locations near Sacramento become rather clean as in SV cluster 3, despite moderately high PC1 scores. SV cluster 5 is characterized by overall high ozone levels both in Sacramento (high PC1 scores in Figure S3 in Supporting Information) and downwind to the north (high PC2 scores in Figure S3). The characteristics of each cluster are summarized in Table 2.

Subregional daily cluster memberships are plotted for the entire simulation period in Figure 8. Similar to the SJV, the highest ozone cluster in the Sacramento Valley does not follow immediately after the cleanest days (clusters 0, 1, and 3), but rather evolves over multiple days. Ozone levels in the SFB respond more rapidly to changes in mesoscale meteorology (e.g., changes in strength of the sea breeze). A clean day in SFB (clusters 0, 1 and 2) can be followed by a high ozone day (clusters 3 and 4), which entails ozone increases of as much as 10 ppb at hot spots in the Livermore and Santa Clara Valleys, according to the mean ozone anomaly maps shown in Figure 6.

4 Discussion

4.1 Interbasin Coupling

The coupling between upwind airbasins and SJV can reveal source-receptor relationships that are associated with different meteorological processes. The joint distribution of ozone clusters in SJV in relation to SFB and SV is presented in Table 3. Moderate high ozone clusters 3 and 5 in the SJV are mostly associated with a relatively clean SFB ozone pattern (cluster 2), while SJV cluster 4 is mostly associated with high-ozone SFB clusters 3 and 4. High ozone levels in SFB (cluster type 4) sometimes occur when SJV experiences its lowest ozone levels (cluster 1), which could be caused by the high ozone sensitivity to SFB meteorology. SFB produces and accumulates ozone over

the course of a day when it experiences a sudden weakening of westerly flows that are insufficient to ventilate pollutants. Since the transport of pollutants from the SFB is diminished, the SJV can maintain lower ozone levels. There are fewer clear relationships between SJV and SV ozone patterns, partly due to smaller sample sizes in the SV. When SV is experiencing its highest ozone levels, SJV ozone levels are generally high (clusters 3-6), and low SJV ozone levels (clusters 1 and 2) do not occur on the highest SV ozone days (cluster 5). Among moderate ozone days for the SJV (clusters 3, 4, and 5), clusters 3 and 5 are associated with moderate to high SV ozone levels, while cluster 4 is associated with relatively low SV ozone levels.

4.2 Effects of Weekend Emission Reductions on Spatial Distribution of ozone

The variations of ozone patterns induced directly (through chemistry and physical processes) and indirectly (through the effects on biogenic emissions) by changes in meteorology can be confounded by changes in anthropogenic emissions. When using ozone spatial patterns in the SJV to reveal meteorological regimes, it is important to understand whether ozone patterns are significantly affected by changes in anthropogenic emissions. Anthropogenic emissions used in this modeling study have day-of-week variations, with lower weekend emissions than on weekdays (see Tonse et al. 2008). To investigate the weekend effect on ozone spatial patterns, we simulated all the weekend days with weekday emissions and examined the variations of ozone spatial patterns induced by the emission change. Ozone differences on Sundays (ΔO_{3wnd}) were used in this analysis.

$$\Delta O_{3wnd} = O_{3a} - O_{3h} \quad (6)$$

where, O_{3a} are the Sunday ozone concentrations simulated using actual Sunday emissions, and O_{3h} is the Sunday ozone concentrations simulated using typical weekday emissions. Saturdays are excluded from the analysis because they can be affected by carryover of Friday emissions.

ΔO_{3wnd} is projected onto the same SJV ozone PCs, 1 to 3, described in Section 3.1. The scores (shown in Figure 9) can be compared to the overall variations in ozone anomalies observed for whole summer (Figure 4). Scores in Figure 12 suggest that the reduction of weekend emissions will reduce overall ozone levels (PC1) and increase the spatial imbalances (PC2 and PC3). Compared to Figure 4, these changes are small relative to the overall variations seen for the whole summer and within each cluster. For example, in Figure 4, the scores on PC1 across all clusters range from –500 to more than 1000; even within each cluster, the scores on PC1 generally vary by more than 300. In contrast, weekend effects only induce average variations of ~100. These minor effects are confirmed by reclustering SJV ozone patterns simulated using all weekday emissions. The new results are not significantly different from the ones we presented previously (see Table 4), in that the tabulation shows only a few off-diagonal elements. The column total represents the "true" clustering results that are driven only by the variations in meteorology. Within each column, distribution among the rows reflects the effects of changes in anthropogenic emissions. The off-diagonal elements in each column can be explained by emission changes occurring on weekends, e.g., in the last column, the overall decreases in weekend ozone concentrations lead to assigning three days in cluster 6 to cluster 5. Similarly, in column 5, three days in cluster 5 are assigned to cluster 2, which has a similar spatial pattern but lower ozone levels. This analysis confirms that the variations seen in ozone spatial patterns are mostly attributable to changes in meteorology.

4.3 Representativeness of ozone measurement sites

Surface measurements have better temporal coverage but may not be spatially representative. The ozone observation sites in Central California are shown in Figure 1 for each of the three subregions. Modeling results on highly resolved grids can be used to investigate whether there are adequate and appropriately located observational sites to reflect region-wide and subregional ozone patterns.

The same clustering procedure is now repeated using modeled data but only at observational sites. Clean days (cluster label 0) here are defined to be those with no more than one site exceeding 84 ppb (8 h maximum ozone). Resulting cluster memberships are compared with those from modeled data in Table 5.

Cluster memberships determined by the two approaches are in reasonable agreement for both the SFB and the SV, since there are few off-diagonal elements. Measurement sites in SJV can distinguish relatively high ozone days (clusters 3, 4, 5, 6) from relatively low ozone days (clusters 1 and 2) and clean days (cluster 0). As for more subtle differences in the clusters, however, SJV observation sites could not distinguish cluster 1 from 2, or cluster 4 from 5 (highlighted in red). About 70% of the days that should be in cluster 4 have been grouped into cluster 5. These clusters are separated by the contrast in ozone between the western and eastern sides of the SJV, which is likely to be caused by different meteorological conditions that yield different ozone production and transport patterns. Figure 1 indicates measurement sites in the SJV tend to be located in more populated areas in the middle or eastern side of the valley. The western side of the SJV is under-sampled. As a result, the current distribution of measurement sites in the SJV alone cannot resolve the spatial contrast between the western and eastern sides of the valley adequately, so that days having ozone distributions characteristic of clusters 4 or 5 tend to be lumped together. Recall that SJV cluster 4 is generally associated with high SFB ozone, while SJV cluster 5 is linked to low SFB ozone. Including SFB sites in the clustering procedure applied to observations can help to distinguish days in SJV ozone cluster 4 from those in cluster 5 without knowing the actual ozone concentrations on the western side of the SJV.

5 Conclusions

In this paper, we applied k-means cluster analysis to daily 8 h ozone maxima modeled for a summer season to characterize meteorology-induced variations in the spatial distribution of ozone air pollution. Principal component analysis is employed to form a reduced dimension set to describe and interpret ozone spatial patterns. The first

three principal components (PCs) capture ~85% of total variance, with PC1 describing a general spatial trend, and PC2 and PC3 each describing a spatial contrast. Results were reported for the San Joaquin Valley, and in relationship to clustering results for two upwind air basins: San Francisco Bay area and Sacramento Valley. Six clusters were determined for the SJV, with two lower ozone clusters 1 and 2, three moderate ozone clusters 3 to 5, and one high-ozone cluster 6. The moderate ozone clusters are distinguished by elevated ozone levels in different parts of the valley: northern, western, and eastern side, respectively. The SJV ozone clusters have stronger coupling with SFB ozone clusters than with the SV.

Variations in ozone spatial distributions induced by anthropogenic emission changes are small relative to overall variations in ozone over the whole summer. Therefore, differences in the ozone regimes characterized here are mostly attributable to direct and indirect meteorological effects. Existing measurement sites are able to capture ozone spatial patterns in the SFB and SV, whereas the western side of the SJV is under-represented.

6 Acknowledgement

The authors thank the Technical Committee overseeing the Central California Ozone Study for helpful comments at many stages of this research. This research was supported by the Assistant Secretary of Fossil Energy, Office of Natural Gas and Petroleum Technology through the National EnergyTechnology Laboratory under the U.S. Department of Energy Contract No. DE-AC02-05CH11231, California Energy Commission, and the Central California Air Quality Study Agency,. The statements and conclusions in this paper are those of the authors, and do not necessarily reflect the views of the sponsoring agencies.

7 References

Anderberg, M. R. (1973). Cluster Analysis for Applications. Monographs and Textbooks on Probability and Mathematical Statistics. Academic Press Inc., New York, 359 pp.

Beaver, S. and Palazoglu, A. (2006). A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. *Atmospheric Environment* **40**(4): 713-725.

Beaver, S., Palazoglu, A., Tanrikulu, S. (2008). Cluster Sequencing to Analyze Synoptic Transitions Affecting Regional Ozone. *Journal of Applied Meteorology And Climatology* **47**(7): 901-916.

Beaver, S., Palazoglu, A. (2009). Influence of synoptic and mesoscale meteorology on ozone pollution potential for San Joaquin Valley of California. Atmospheric Environment **43**: 1779-1788.

Bell, M. L. and Dominici, F. (2008). Effect Modification by Community Characteristics on the Short-term Effects of Ozone Exposure and Mortality in 98 US Communities. *Am. J. Epidemiol.* **167**(8): 986-997.

Dayan, U., and I. Levy (2002). Relationship between synoptic-scale atmospheric circulation and ozone concentrations over Israel, J. Geophys. Res., 107(D24), 4813, doi:10.1029/2002JD002147.

Darby, L. S. (2005). Cluster analysis of surface winds in Houston, Texas, and the impact of wind patterns on ozone. *Journal of Applied Meteorology* **44**(12): 1788-1806.

Dawson, J. P., Adams, P. J. and Pandis, S. N. (2007). Sensitivity of ozone to summertime climate in the eastern USA: A modeling case study. *Atmospheric Environment* **41**(7): 1494.

Emberson, L. D., Bder, P., Ashmore, M. R., Mills, G., Jackson, L. S., Agrawal, M., Atikuzzaman, M. D., Cinderby, S., Engardt, M., Jamir, C., Kobayashi, K., Oanh, N. T. K., Quadir, Q. F. and Wahid, A. (2009). A comparison of North American and Asian exposure-response data for ozone effects on crop yields. *Atmospheric Environment* **43**(12): 1945.

Everitt, B. S. (1993). Cluster Analysis, 3rd ed. London, Heinemann Education.

Everitt, B. S., Landau, S. and Leese, M. (2001). Cluster Analysis. 4th ed. Edward Arnold Publishers, 229 pp.

Fujita, E., Keislar, R., Stockwell, W., Moosuller, H., DuBois, D., Koracin, D. and Zielinska, B. Central California Ozone Study-Volume I, Field Study Plan. Division of Atmospheric Science Desert Research Institue, 2215 Raggio Parkway, Reno, NV. 1999

Hotelling, H. (1935). The most predictable criterion. *J. Ed. Psych.* **26**: 139-142.

Jin, L.; Brown, N.J.; Harley, R.A.; Bao, J.; Michelson, S.A.; Wilczak, J.M. (2010). Seasonal versus Episodic Performance Evaluation for an Eulerian Photochemical Air Quality Model. *Journal of Geophysical Research* 115, D09302, doi: 10.1029/2009JD012680.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, Wiley.

Kaufmann, P. and Weber, R. O. (1996). Classification of Mesoscale Wind Fields in the MISTRAL Field Experiment. *Journal of Applied Meteorology* **35**(11): 1963-1979.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. Technical report, Statistical Forecast Project Report 1, Dept. of Meteor. MIT, 49 pp.

Ludwig, F. L., Jiang, J.-Y. and Chen, J. (1995). Classification of ozone and weather patterns associated with high ozone concentrations in the san francisco and monterey bay areas. *Atmospheric Environment* **29**(21): 2915.

Pearson, K. (1902). On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **2**: 559-572.

Tao, Z., Williams, A., Huang, H.-C., Caughey, M. and Liang, X.-Z. (2007). Sensitivity of U.S. surface ozone to future emissions and climate changes. *Geophys. Res. Lett.* **34**.

Tonse, S.; Brown, N.J.; Jin, L.; Harley, R.A. (2008). A Process-Analysis Based Study of the Ozone Weekend Effect. *Atmospheric Environment* 42, 7728-7736.

Table 1 Description of the first three PCs for each subregion.

Subregion	Principal Component	Descriptions
SJV	1	Largely positive, describing a general spatial pattern in ozone concentrations (higher loadings seen along Highway 99, and near Fresno and Bakersfield).
	2	A spatial contrast in southeastern direction, with positive loadings in the Sierras and Mojave desert, and negative loadings in the northwest (Modesto to Fresno).
	3	A spatial contrast in northeastern direction, with positive loadings in the Sierras and northeastern SJV, and negative loadings along the southwestern side of the valley.
SFB	1	Largely positive, describing a general spatial pattern in ozone concentrations (larger loadings seen in Livermore valley and south bay).
	2	An east-west spatial contrast between coastal and inland areas, with positive loadings on the coastal side.
	3	A north-south spatial contrast with positive loadings on the northern side.
SV	1	Largely positive, describing a general spatial pattern in ozone concentrations (larger loadings near Sacramento).
	2	A north-south spatial contrast with positive loadings in northern, more rural areas
	3	An east-west spatial contrast between the Sacramento Valley and Mountain County air basins, with positive loadings in the Sierras.

Table 2 Descriptions of ozone clusters

Subregion	Ozone level	Spatial pattern cluster	Descriptions of ozone anomalies
SFB	Low	1	Mostly negative, indicating lower than average ozone levels
		2	Very small ozone increases on the eastern side (more to the inland side) of the basin
	High	3	Elevated ozone concentrated on the southeastern side of the basin
		4	Elevated ozone levels on the western, central, and southern side (more to the ocean side) of the basin
SJV	Low	1	Overall lower than average ozone levels, with very small increases on the western side of the valley
		2	Small ozone increases in the Sierras and the Mojave desert, downwind side of the major urban areas.
	Moderate	3	Moderate ozone increases on the north side
		4	Moderate ozone increases on the western side
		5	Moderate increases on the southeastern
	High	6	Strong ozone increases throughout the valley
SV	Low	1	Very small ozone increases in the Sacramento area (southern end of the valley)
		2	Small ozone increases near Sacramento
		3	Strong ozone increases in the north, but decrease in the Sacramento, resulting in rather clean conditions in the usual ozone hot spots.
	Moderate	4	Moderate ozone increases both in the south and southeast Sacramento area
	High	5	Strong ozone increases throughout the valley

Table 3 Inter-basin relationships of ozone clusters.

SFB cluster	SJV cluster 1	2	3	4	5	6
1	2	17	3	0	3	0
2	2	4	11	1	16	3
3	0	1	2	6	0	6
4	5	0	1	6	0	1

SV cluster	SJV cluster 1	2	3	4	5	6
1	5	3	2	4	0	1
2	1	3	4	4	6	2
3	2	10	2	0	2	0
4	0	3	4	2	6	3
5	0	0	4	1	3	1

Each entry at i^{th} row and j^{th} column is the number of days that has ozone cluster types i and j in the respective subregions.

Table 4 Comparison of SJV clustering results with all weekday emissions only versus day-of-week dependent emissions[4]

SJV clusters (c1~c6)	Meteorology induced variation
Meteorology + Emission	

	c1	c2	c3	c4	c5	c6
c1	8	0	0	0	0	0
c2	0	15	3	0	4	0
c3	0	1	14	1	0	0
c4	1	0	0	12	0	0
c5	0	0	4	0	12	3
c6	0	0	1	0	0	9

[4] An outlier day 265 is removed from the clustering due to its exceedingly large score on PC 3.

Table 5 Comparison of clustering results with modeled data at all grids versus using modeled data only at observational sites.

SJV

grids \ sites	0	1	2	3	4	5	6
0	*25*	2	0	0	0	0	0
1	3	*4*	0	2	0	0	0
2	1	9	*8*	4	0	0	0
3	0	0	0	*11*	1	5	0
4	0	0	0	1	*3*	8	1
5	0	0	2	0	0	*17*	0
6	0	0	0	0	1	1	*8*

SFB

grids \ sites	0	1	2	3	4
0	*25*	0	2	0	0
1	3	*15*	7	0	0
2	1	2	*32*	2	0
3	0	0	1	*12*	2
4	0	0	0	0	*13*

SV

grids \ sites	0	1	2	3	4	5
0	*24*	0	1	1	0	0
1	4	*10*	1	0	0	0
2	0	4	*16*	0	0	0
3	0	1	1	*14*	0	0
4	0	0	2	0	*12*	4
5	0	0	0	1	0	*8*

The bold faced numbers are cluster labels (row labels are clusters resulting from using modeled data at all grids, column labels are clusters resulting from using modeled data at measurement sites).

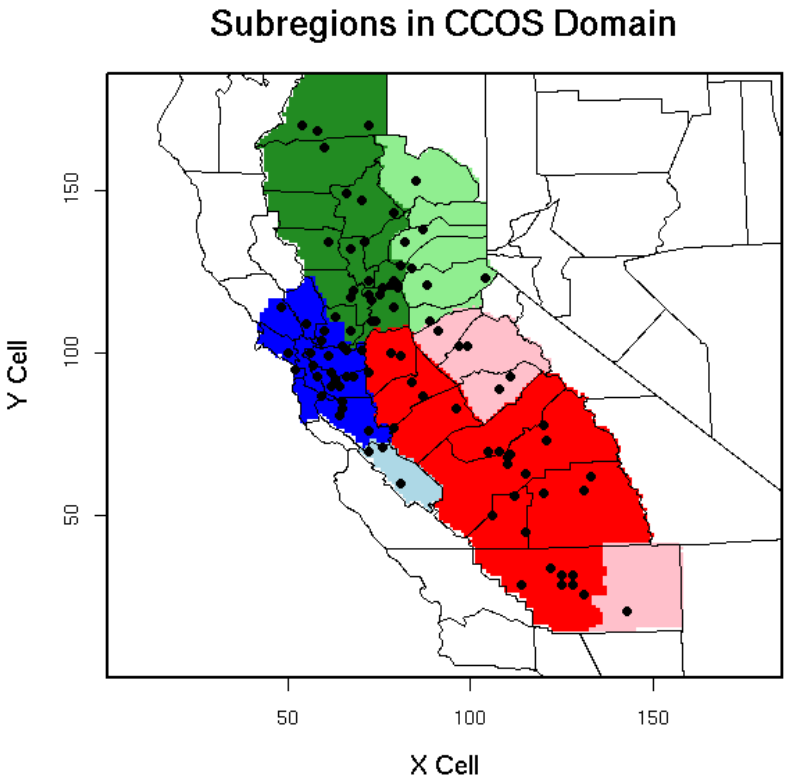

Figure 1 Subregions in Central California used for clustering ozone spatial patterns. Black dots are actual ozone measurement sites. SJV (red+pink), SFB (blue+lightblue), SV (green + lightgreen).

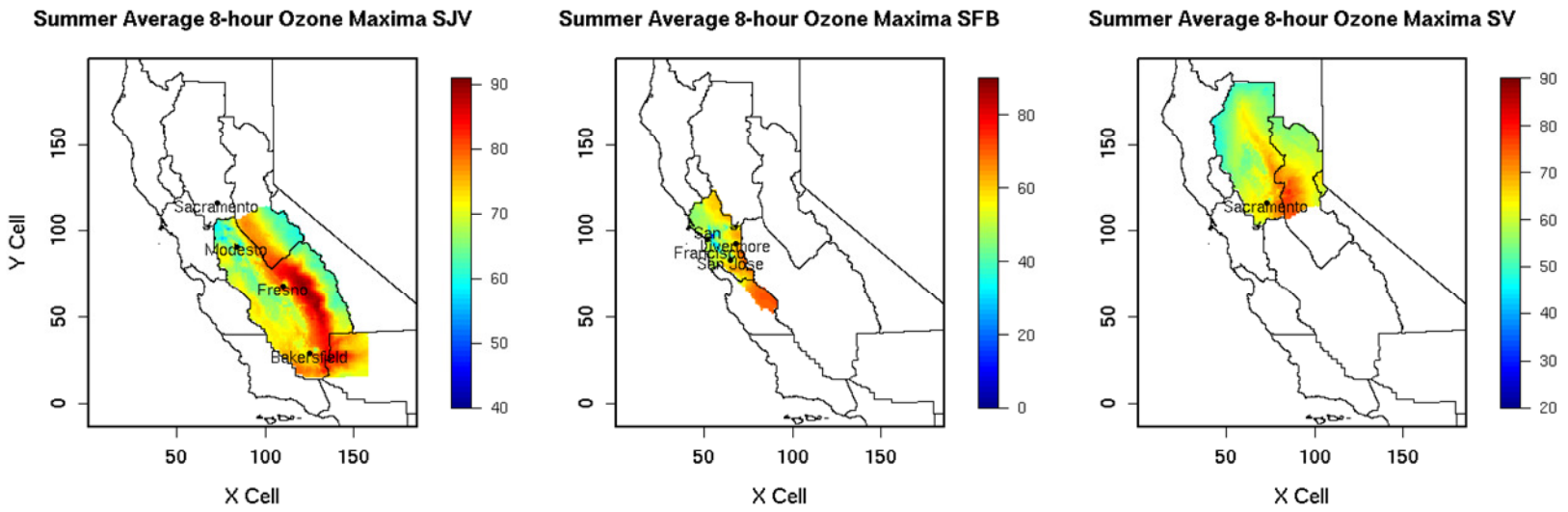

Figure 2 Summer average 8 h ozone maxima (ppb) for the three subregions.

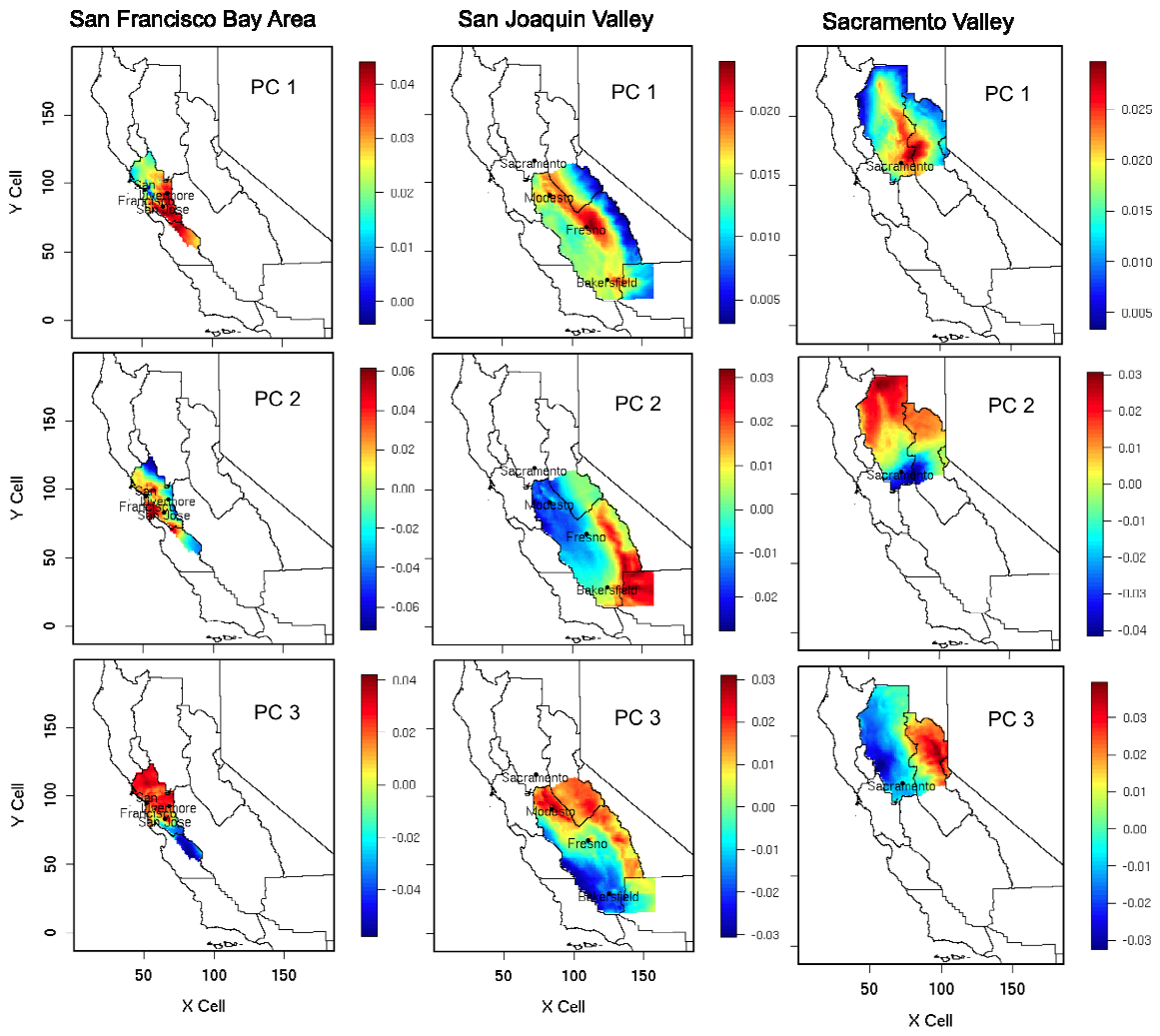

Figure 3 Ozone loadings (ppb) on the three PCs of 8 h ozone maxima.

Figure 4 Visualizing SJV ozone clusters (1 to 6) with PC scores.

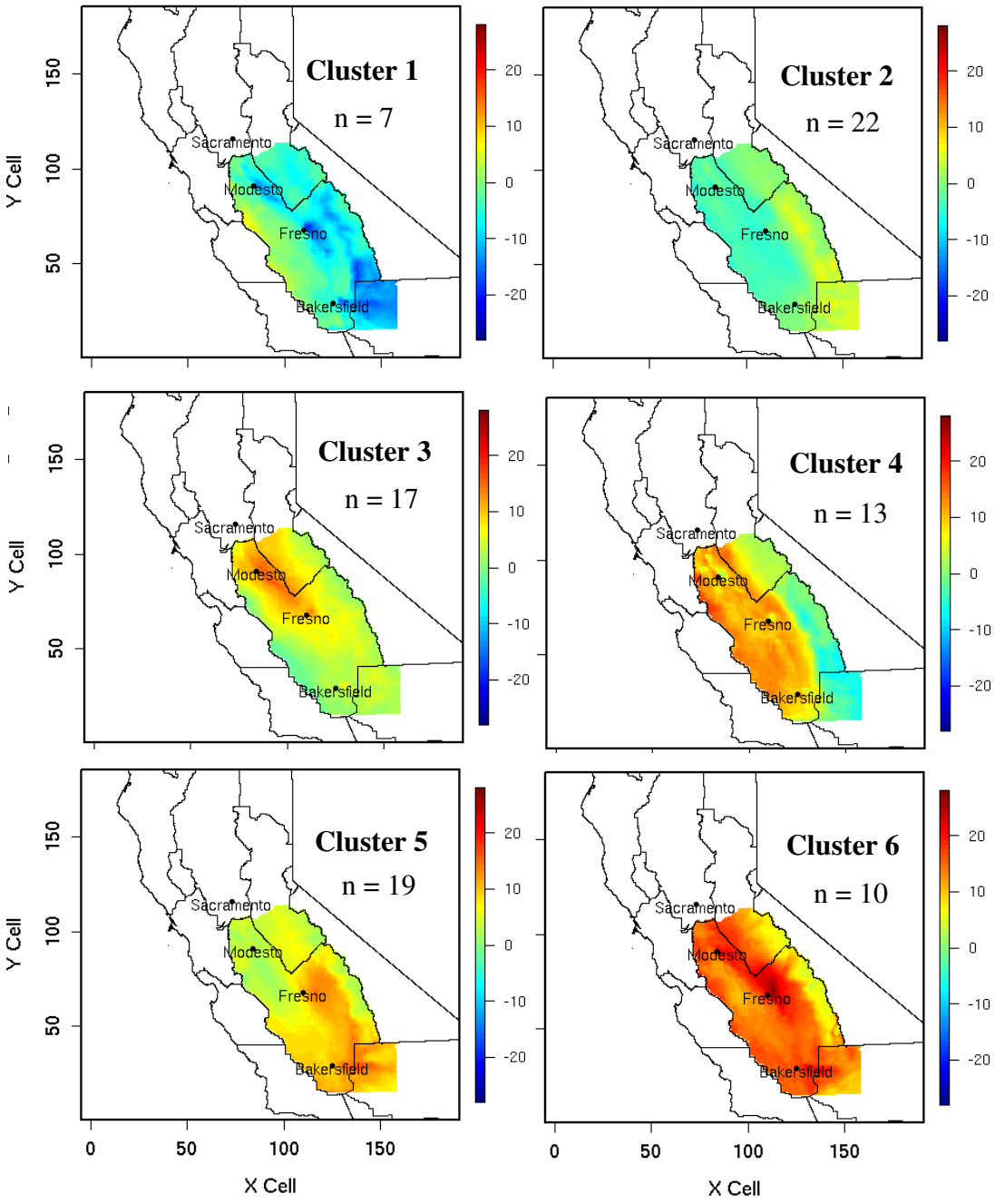

Figure 5 Ozone anomaly field (ppb) averaged over SJV clusters.

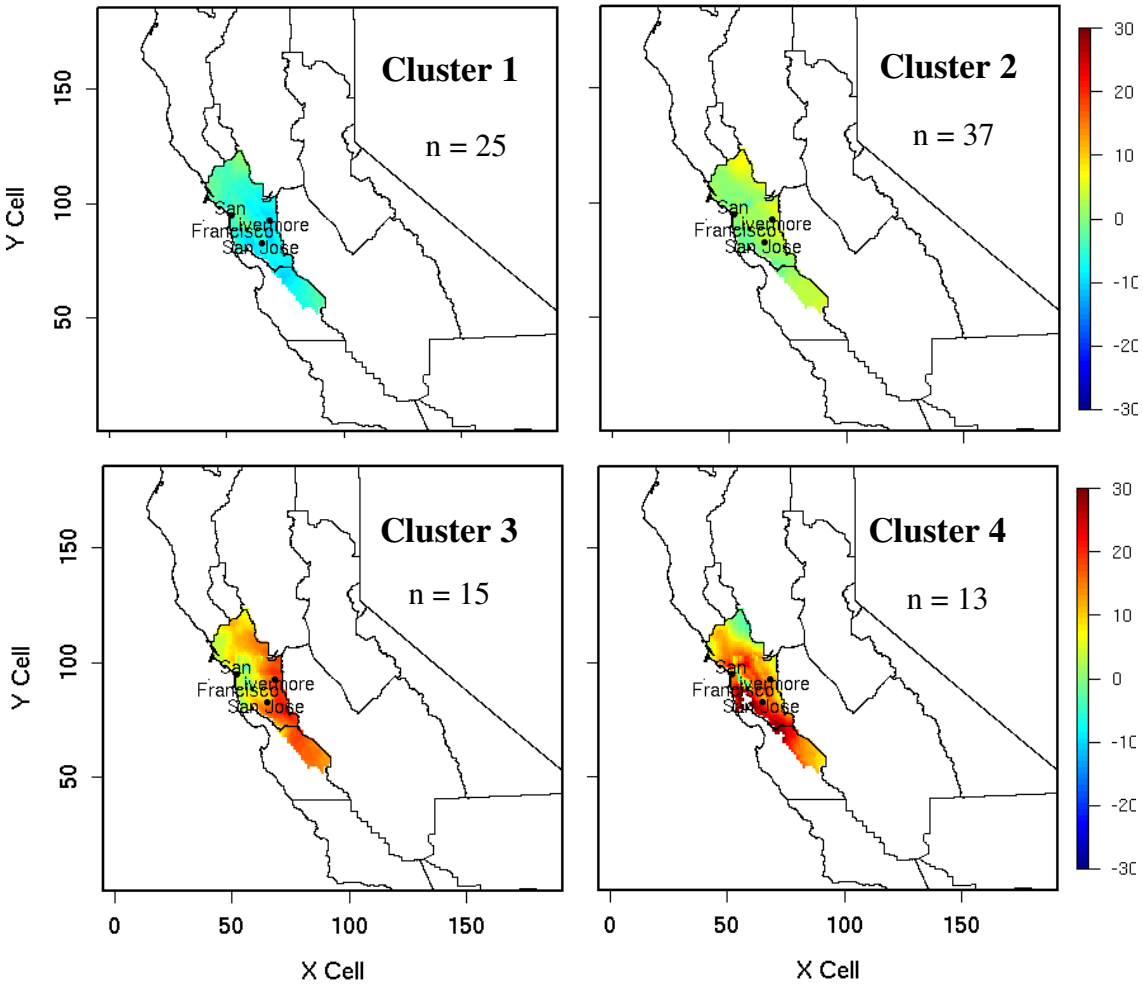

Figure 6 Mean ozone anomaly fields (ppb) patterns of SFB clusters.

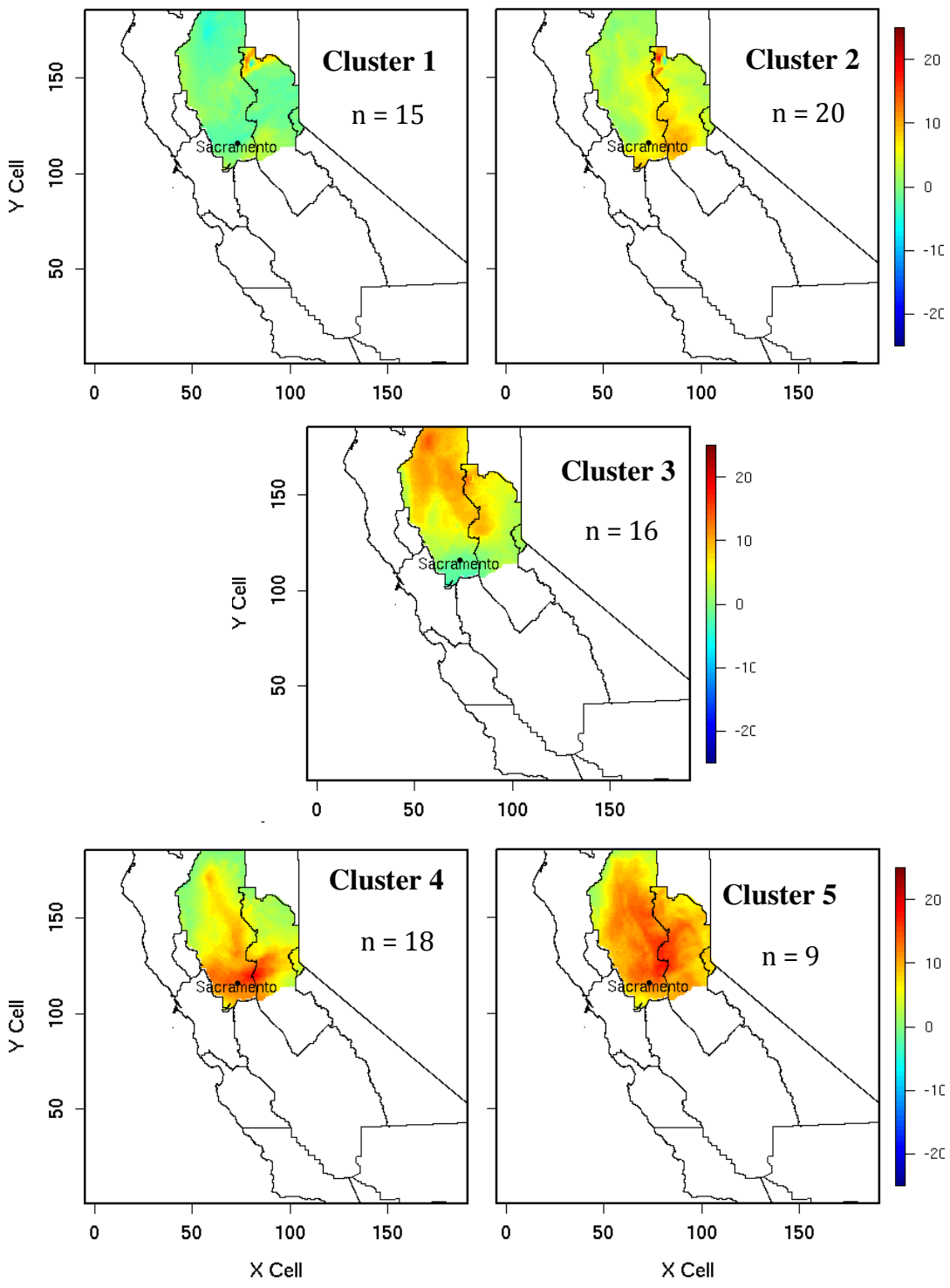

Figure 7 Mean ozone anomaly fields (ppb) patterns of SV clusters.

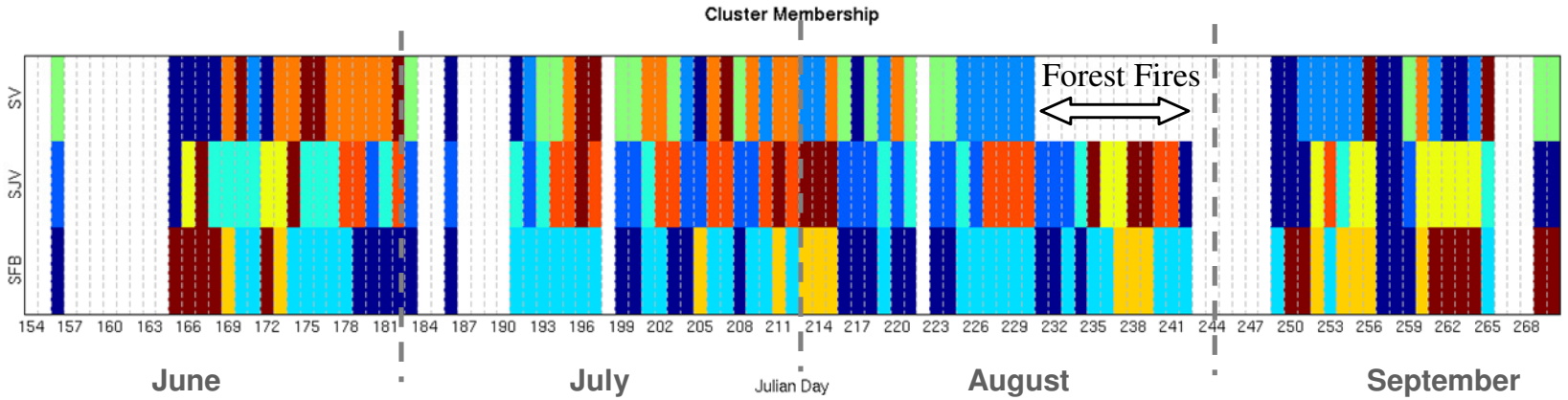

Figure 8 Cluster memberships.

Clean days (cluster 0) are blank. Days with significant forest fire influence on SV ozone levels are indicated on the plot.

SJV Key (cluster type from 1 to 6)
SFB Key (cluster type from 1 to 4)
SV Key (cluster type from 1 to 5)

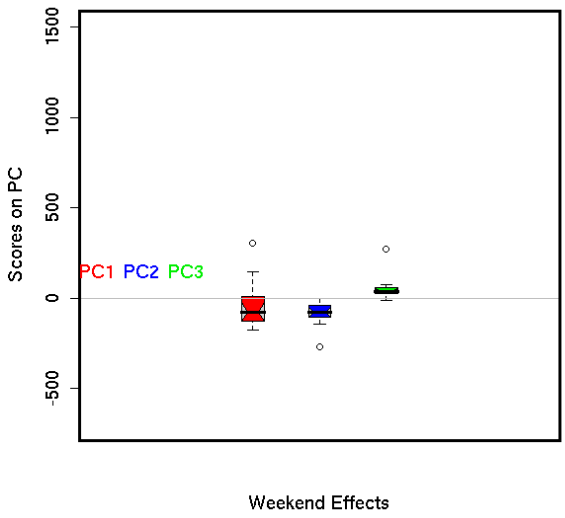

Figure 9 SJV weekend effects on ozone spatial variations projected onto the first three PCs.