6.4

AGGREGATION USING CLUSTER ANALYSES FOR MODELS-3 CMAQ APPLICATIONS

Sharon LeDuc*, Brian Eder*, and Robin Dennis*
Atmospheric Sciences Modeling Division, Air Resources Laboratory,
National Oceanic and Atmospheric Administration, Research Triangle Park, NC 27711

Richard Cohn
Analytical Sciences, Inc.
Durham, NC 27713

1. INTRODUCTION

Models-3, a framework for air quality modeling, is scheduled for public release in June 1998. The framework will support statistical analysis and presently uses four SAS® modules (base, graph, AF and FSP) in the emissions processing. Statistical capability in Models-3 is also found in PAVE (Thorpe , 1996) which displays time series plots. Statistics such as mean, median, and percentiles can be plotted from Models-3 with IBM's Visualization Data Explorer .

The statistical tool described in this paper is for policy planning for air quality issues related to annual or multi-year measures, rather than the most extreme events. Scientifically credible and reliable estimates of air quality for large regions rely on air quality models, such as the Community Multiscale Air Quality (CMAQ) model in Models-3. Application of such models requires massive resources, both human and computer, for each policy and/or meteorological scenario. Analysis of benefits proposed for the Clean Air Act Amendments of 1990 requires annual timescales. Unfortunately, CMAQ model, like most Eulerian models, challenges the practical limits of current computer resources as well as our ability to collect the pertinent input data on annual scales. As a result, applications to determine the long-term relationship between changing emissions patterns and ambient air concentrations are resource intensive.

To circumvent this problem, a statistical aggregation method, initially developed for RADM acid-deposition applications (Brook et al., 1995), will be modified to provide estimates of long-term (seasonal or annual) ambient air concentrations, wet and dry deposition amounts, and measures related to visibility. An important feature of the aggregation method is to represent source attribution. The method uses cluster analysis based on the premise that, at any given location, ambient air concentrations (also deposition amounts) can be represented by a finite number of different, though recurring, meteorological regimes. The meteorological regimes identified for RADM considered only the eastern U.S. and Canada where acid deposition was the issue. Now air quality issues, such as regional haze, are more geographically extensive. The sample selection and aggregation weightings used with RADM need to be reexamined for use with these issues.

Corresponding author address: Sharon LeDuc (MD-80). On assignment to National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711; email: leduc@hpcc.epa.gov

2. DATA

2.1 Meteorological

To accommodate the continental domain and to achieve sufficient spatial resolution, the cluster analysis uses data at 336 grid nodes with 2.5° spatial resolution from the NCEP/NCAR 40-year reanalysis project (Kalnay *et al.*, 1996). The RADM clusters were defined using 850mb winds, but because of the mountains in the western U.S. the 700 mb wind components for 1800 UTC were used here. The domain is designed to prevent excessive influence from ocean-based meteorology. Since a model (RADM or CMAQ) is usually run for a 5-day period (the first two days establish initial conditions with model predictions from days 3-5 saved as a "3-day event"), 5-day periods were clustered instead of 3-day periods used for RADM.

2.2 Air Quality

Assignment of a 5-day period to a cluster, will determine how air quality data, model estimates or monitoring measures, for that period will be used to estimate annual statistics. Evaluating how well the meteorologically derived clusters relate to air quality requires air quality monitoring data for the same period of record as the meteorological data. Air quality data are more limited than the meteorological data. Surrogate air quality data, derived from human observations of visible range at airports, will be used first. The near noon observation, converted to an extinction coefficient (Husar and Wilson, 1993) has an

inverse relationship with fine particles in the air. Later, other sources of air quality data will be used for shorter time periods: National Atmospheric Deposition Program (NADP); Clean Air Status and Trends Network (CASTNet); Interagency Monitoring of Protected Visual Environments (IMPROVE); and Aerometric Information Retrieval System (AIRS).

3. METHODOLOGY

3.1 Clustering

The purpose of objectively defining meteorological categories is to identify recurring atmospheric transport patterns associated with varying concentration and deposition patterns of air pollutants. Identification of these patterns facilitates selection of time periods, i.e. the sample, for simulation by CMAQ. Model output from the sample will be weighted in the aggregation based on population weights of the clusters or strata. Representative meteorological categories have been explored by others (Fernau and Samson, 1990; Davis and Kalkstein, 1990). The approach used here is based on a variation of the methods previously used by Brook *et al. (op cit)* in selecting a RADM 30-episode sample for aggregation.

The common thread is the cluster analysis of zonal *u* and meridional *v* wind components. A 10-year period (1980-1990) was used. To make the analysis computationally feasible, the first, third, and fifth days of each 5-day episode were considered. Clusters were initially defined based upon "consecutive" rather than "running" or overlapping 5-day periods from 1980-1985. Then, each remaining episode ("running" 5-day periods from 1980 through 1990) was classified into the cluster that minimizes the sum (over the 336 grid nodes and three days) of the squared deviations of each *u* and *v*. Cluster analyses were carried out using SAS®. However, due to the extreme computational burden of these analyses, it was necessary to calculate the distance matrix externally from the clustering procedure itself.

Winds can be defined in polar coordinates as well as in the Euclidean *u* and *v* coordinate system. Preliminary cluster analyses investigated polar coordinate systems, clustering with the angle (direction) defined by the wind vector. Results suggested that clusters defined with polar coordinates didn't discriminate seasonal differences in wind vector patterns as well as clusters defined using Euclidean coordinates. Later analyses with *u* and *v* coordinate investigated four clustering variations:

- 30 clusters, defined using annual data (consecutive 5-day periods from 1980-1985, as described above)
- 30 clusters, defined seasonally (15 clusters defined from the warm season Apr.-Sept. period, and 15 clusters defined from the cold season Oct.-Mar. period)
- 60 clusters, defined using annual data
- 60 clusters, defined seasonally (30 clusters defined from the warm season Apr.-Sept. period, and 30 clusters defined from the cold season Oct.-Mar. period)

For the annual defined clusters, remaining 5-day events (running 5-day periods from 1980-1990) were then classified based on minimizing the squared distance from the cluster average. For the seasonally defined clusters, the remaining 5-day events were classified into clusters defined for the same season as the event. For the 30-cluster analyses, cluster averages were displayed as dots on maps illustrating the intracluster variability in the wind fields. Star chart histograms were used to illustrate the frequency of occurrence of events from each cluster, for each month of the year.

3.2 Aggregation

The aggregation approach is based upon weights determined for meteorological categories that account for a significant proportion of the variability. These categories need to account for variation in the air quality measures as well. Within and between cluster variability of air quality measures will be examined. The extinction coefficient will be used first as was done for RADM (Eder *et al.*, 1996). Other air quality characterizations will be evaluated with available air quality data sets. Air quality is feature of interest, but what is considered when weights are based on strata of wind, are transport mechanisms involved in the associated atmospheric processes, and in particular that source-attribution analyses be facilitated. This requires that clusters reflect wind flow parameters. Other parameters are important, but may not be necessary in defining strata or clusters since wind field patterns in essence describe frontal passages, along with their meteorological properties. Evaluation of aggregation results may require additional parameters in the defining of clusters.

4. RESULTS AND FUTURE WORK

Maps of mean wind vectors were done for each of the 30 clusters defined using annual data. The clusters are ordered according to overall frequency of occurrence, with Cluster 1 being most prevalent and Cluster 30 least prevalent. Mean vectors for day 5 of the 5-day events of Cluster 11 (Fig.1a) illustrate average behavior associated for a cluster, but do not indicate the variability inherent in the cluster. For example (Figure 1b) illustrates the wind field for the fifth day of an individual event (Dec. 19-23, 1989) that was assigned to Cluster 11 (Fig.1a). Compared to the mean wind field for day 5 of Cluster 11 reveals fairly close resemblance between the two. By contrast, (Fig.1c) depicts the fifth day from another event (Dec. 3-7, 1990) belonging to Cluster 11. This pattern does

not resemble the mean wind field nearly as well.

Simultaneously viewing all of the wind fields assigned to the fifth day of Cluster 11 (Fig.1d) shows the mean wind vectors for day 5 of Cluster 11 on a thinned-out grid that only includes alternating grid nodes. Each group of dots depicts the location of the wind vector arrowheads for individual events assigned to this cluster and collectively illustrate the distribution of arrowheads for all events belonging to Cluster 11.

The star chart histograms of the 30 clusters defined using annual data (Figure 2) illustrate the frequency of occurrence of 5-day events belonging to several clusters. Events from Cluster 1 accounted for 13.88% of all 5-day events between 1980 and 1990, those from Cluster 8 accounted for 3.76% and Cluster 11, 3.29%. The numbers arranged radially on each chart depict the number of events belonging to the cluster from each month of the year. The length of the line pointing to each month is proportional to this frequency of occurrence, and the ends of the lines are connected to facilitate the visualization of patterns. Several observations may be made:

- Although defined using annual data, the cluster frequencies reveal definite seasonal tendencies. Clusters do not occur randomly throughout the year, but rather exhibit a tendency to occur more frequently within specific seasons. The clustering procedure successfully identifies and discriminates wind field patterns that are associated with seasonally distinct meteorological classes.
- While clusters containing summer events tend to be quite distinct from those containing winter events (and vice versa), many clusters contain events from a combined "transitional" season that includes spring and fall months.
- The three most prevalent clusters heavily emphasize the summer months; however, these months are rarely represented by the remaining 27 clusters (not shown).

The disportionate representation of non-summer events in the set of 30 clusters is not surprising, since the wind fields are expected to be less variable in the summer. Seasonal differences in meteorology and atmospheric chemistry are important in explaining the variability exhibited by the air quality parameters of interest. Evaluating the relationship of these clusters to air quality is still in progress.

5. REFERENCES

Brook, J.R, Samson, P.J., and Sillman, S., 1995: Aggregation of selected three-day periods to estimate annual and seasonal wet deposition total for sulfate, nitrate and acidity. Part I: A synoptic and chemical climatology for Eastern North America. *J. Applied. Meteor.* **34**, 297-325.

Brook, J.R.; Samson, P.J.; and Sillman, S., 1995: Aggregation of selected three-day periods to estimate annual and seasonal wet deposition total for sulfate, nitrate and acidity Part II: Selection of events, deposition totals and source-receptor relationships. *J. Appl d. Meteor.* **34**, 326-339.

Davis, R. E. and Kalkstein, L. S., 1990: Development of an automated spatial synoptic climatological classification. *International Journal of Climatology* **10**, 769-794.

Eder, B.K. and LeDuc, S.K., 1996: Can selected RADM simulations be aggregated to estimate annual concentrations of fine particulate matter. Preprints of the *11th Annual International Symposium on the Measurement of Toxics and Related Air Pollutants,* RTP, NC, pp. 732-739.

Eder, B.K. and LeDuc, S.K. and F.D.Vestal., 1996: Aggregation of selected RADM simulations to estimate annual ambient air concentrations of fine particulate matter. Preprints of AMS *9th Joint Conference on the Application of Air Pollution Meteorology with AWMA*, Jan. 28-Feb. 2, 1996, Atlanta, GA, pp. 390-392.

Fernau, M.E. and Samson, P.J., 1990: Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: Transport patterns. *J. of Applied Meteor.* **29**, 735-750.

Husar, R.B. and W.E. Wilson, 1993. Haze and sulfur emission trends in the eastern U.S.. *Environ. Sci. Technol.*. 27, 13-16.

Kalnay, E., M.Kanamitsu, R.Kistler, W.Collins, D. Deaven, L.Gandin, M.Iredell, S. Saha, G.White,J.Woollen, Y.Zhu,M.Chelliah, W.Ebisuzaki,W. Higgins,J.Janowiak, K.C.Mo,C. Ropelewski, J.Wang, A .Leetmaa,.R.Reynolds,R.Jenne,and D.Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**, 437-471.

Thorpe, S., D.Hwang,W.T.Smith,T.L.Turner, 1996. The Package for Analysis and Visualization of Environmental Data. *Proc. of Computing in Environmental Resource Management*,, 2-4 Dec., RTP, NC, AWMA 241-249.

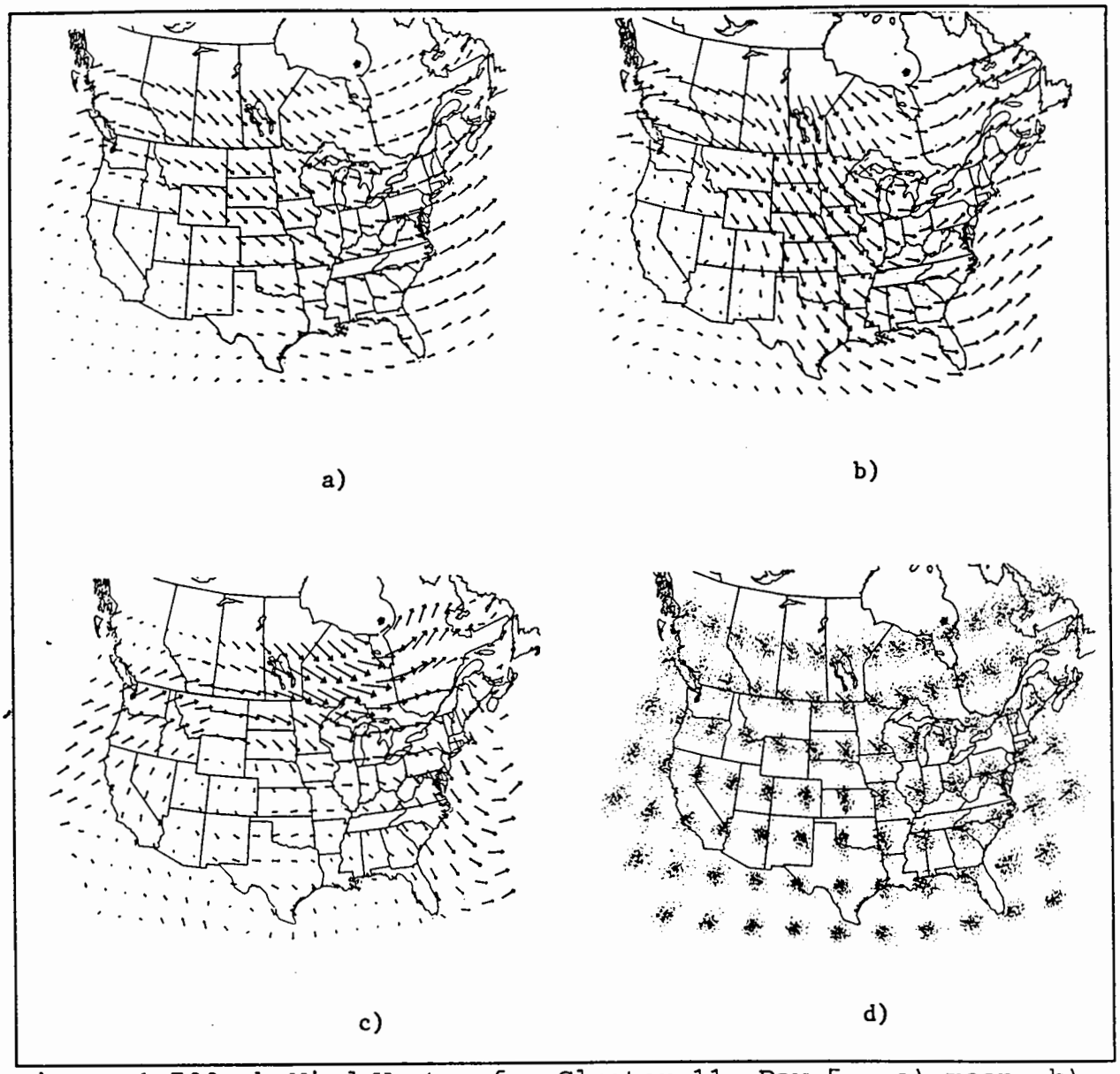

Figure 1 700 mb Wind Vector for Cluster 11, Day 5: a) mean; b) Dec. 23, 1989; c) Dec. 7, 1990; d) variability

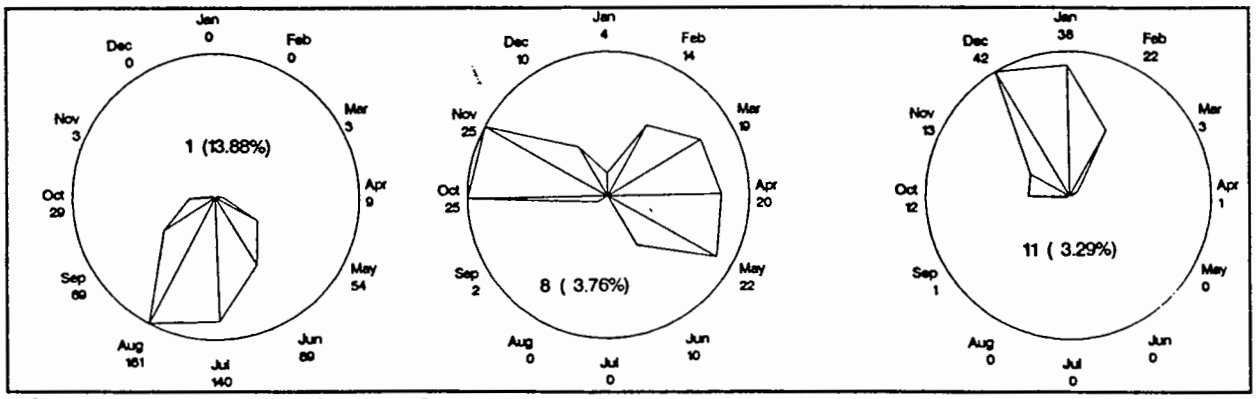

Figure 2 Star chart for Clusters 1, 8, 11

TECHNICAL REPORT DATA

1. REPORT NO.	2.	3.
EPA/600/A-97/081		

4. TITLE AND SUBTITLE	5.REPORT DATE
Aggregation Using Cluster Analyses for Models-3 CMAQ Applications	6.PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8.PERFORMING ORGANIZATION REPORT NO.
Sharon LeDuc, et al. (See title page)	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10.PROGRAM ELEMENT NO.
Atmospheric Modeling Division National Exposure Research Laboratory U.S. Environmental Protection Agency Research Triangle Park, NC 27711	11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS	13.TYPE OF REPORT AND PERIOD COVERED
NATIONAL EXPOSURE RESEARCH LABORATORY OFFICE OF RESEARCH AND DEVELOPMENT U.S. ENVIRONMENTAL PROTECTION AGENCY RESEARCH TRIANGLE PARK, NC 27711	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Models-3, a framework for air quality modeling, is scheduled for public release in June 1998. The framework will support statistical analysis. and presently uses four SAS® modules (base, graph, AF and FSP) in the emissions processing. Statistical capability in Models-3 is also found in PAVE (Thorpe , 1996) which displays time series plots. Statistics such as mean, median, and percentiles can be plotted from Models-3 with IBM's Visualization Data Explorer .

The statistical tool described in this paper is for policy planning for air quality issues related to annual or multi-year measures. rather than the most extreme events. Scientifically credible and reliable estimates of air quality for large regions rely on air quality models, such as the Community Multiscale Air Quality (CMAQ) model in Models-3. Application of such models requires massive resources, both human and computer, for each policy and/or meteorological scenario. Analysis of benefits proposed for the Clean Air Act Amendments of 1990 requires annual timescales. Unfortunately, CMAQ model, like most Eulerian models, challenges the practical limits of current computer resources as well as our ability to collect the pertinent input data on annual scales. As a result, applications to determine the long-term relationship between changing emissions patterns and ambient air concentrations are resource intensive.

To circumvent this problem, a statistical aggregation method, initially developed for RADM acid-deposition applications (Brook et al., 1995), will be modified to provide estimates of long-term (seasonal or annual) ambient air concentrations, wet and dry deposition amounts, and measures related to visibility. The aggregation methods, which use cluster analysis, are based on the premise that, at any given location, ambient air concentrations (also deposition amounts)
are governed by a finite number of different, though recurring, meteorological regimes. The meteorological regimes identified for RADM considered only the eastern U.S. and Canada where acid deposition was the issue. Now air quality issues, such as regional haze, are more geographically extensive. The sample selection and aggregation weightings used with RADM need to be reexamined for use with these issues.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b.IDENTIFIERS/ OPEN ENDED TERMS	c.COSATI

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21.NO. OF PAGES
Release to Public	Unclassified	
	20. SECURITY CLASS *(This Page)*	22. PRICE
	Unclassified	