Chapter 17

AN AGGREGATION AND EPISODE SELECTION SCHEME DESIGNED TO SUPPORT MODELS-3 CMAQ

Richard D. Cohn
Analytical Sciences, Inc.
Durham, NC 27713

Brian K. Eder[*] and Sharon K. LeDuc[**]
Atmospheric Modeling Division
National Exposure Research Laboratory
Research Triangle Park, NC 27711

ABSTRACT

In support of studies mandated by the 1990 Clean Air Act Amendments, the Models-3 Community Multiscale Air Quality (CMAQ) model can be used to estimate pollutant concentrations and deposition associated with specified emission levels. Assessment studies require CMAQ-based distributional estimates of ozone, acidic deposition, $PM_{2.5}$, and visibility on seasonal and annual time frames. Because it is not financially feasible to execute CMAQ over such extended time periods, CMAQ must be executed for a finite number of episodes under a variety of meteorological classes. A statistical procedure called aggregation, must then be applied to the CMAQ outputs to derive seasonal and annual estimates.

The objective of this research is to develop an aggregation approach and set of episodes that would support model-based distributional estimates (over the continental domain) of air quality parameters. The approach utilized cluster analysis and the 700 mb *u* and *v* wind field components over the time period 1984-1992 to define homogeneous meteorological clusters. A total of 20 clusters (five per season) were identified by the technique. A stratified sample of 40 events was selected from the clusters, using a systematic sampling technique.

This stratified sample is then evaluated through a comparison of aggregated estimates of the mean extinction coefficients (b_{ext}) to the actual mean b_{ext} observed at 201 stations nationwide for a nine year period (1984-1992). The b_{ext}, a measure of visibility, was selected for use in the evaluation for two reasons. First, of all of the air quality parameters simulated by CMAQ, this visibility parameter provides one of the most spatially and temporally comprehensive data sets available, and second, the b_{ext} can serve as a surrogate for $PM_{2.5}$ for which little data exists. Results from the evaluation revealed a high level of agreement ($r^2 = 0.988$) indicating that the aggregation and episode selection scheme was indeed representative.

---

[*]On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. Corresponding author address: Brian Eder, MD-80, Research Triangle Park, NC 27711. E-mail: eder@hpcc.epa.gov

[**]On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

17.0 AN AGGREGATION AND EPISODE SELECTION SCHEME DESIGNED TO SUPPORT MODELS-3 CMAQ

17.1 Introduction

This chapter describes the development of an aggregation and episode selection scheme to support the derivation of annual and seasonal estimates of air quality parameters in Models-3 CMAQ. Relevant background information regarding this activity, as well as a precise statement of the present objectives, is provided in this section. Section 17.2 summarizes the important elements of the approach and the overall strategy, including the rationale behind the methods and the limitations associated with them. Details of the development of an aggregation and episode selection are provided in sections 17.3, 17.4, and 17.5. Section 17.6 contains an example application and evaluation of the methodology using one particular air quality parameter (b_{ext}), and section 17.7 provides a summary and discussion.

17.1.1 Background

In support of studies mandated by the 1990 Clean Air Act Amendments, the Models-3 Community Multiscale Air Quality (CMAQ) model is used by EPA Program Offices to estimate deposition and air concentrations associated with specified levels of emissions. Assessment studies and effects models require CMAQ-based distributional estimates (such as annual and seasonal averages) of ozone, acidic deposition, and measures related to visibility. Such estimates would ideally be obtained by using CMAQ to simulate atmospheric chemical processes associated with meteorological conditions occurring on a daily basis over several years. However, for logistical and cost reasons it is not currently feasible to execute CMAQ over an extended time period such as a full year. Therefore, in practice CMAQ must be executed for a finite number of episodes or "events," which are selected to represent a variety of meteorological classes. A statistical procedure called aggregation, must then be applied to the outputs from CMAQ to derive the required annual- and seasonal-average estimates from this finite number of events. The objective of the research described in this chapter is to develop such an aggregation approach and evaluate its effectiveness using the b_{ext}.

The basic problem of developing representative meteorological categories has been explored by other researchers for a variety of purposes, including Fernau and Samson (1990a,b); Davis and Kalkstein (1990); Eder et al. (1994). The approach used here is based on a variation of the methods previously used by Brook et al. (1995a,b) in selecting a 30-event aggregation set for the Regional Acid Deposition Model (RADM). The approach of Brook et al. involved four major components. Cluster analysis of wind fields was used to determine meteorologically representative categories. The determination of the number of clusters to retain was based upon within-group variance patterns and prior work by Fernau and Samson (1990a,b). A procedure for aggregating the episodic results into annual totals and averages involved frequency-weighted sums and estimated deposition-precipitation relationships. Event selection procedures were designed to emphasize categories that accounted for most of the annual wet sulfate deposition,

while also representing some winter and dry events. A summary of some specific elements of their approach is provided in Table 17-1.

Table 17-1 Summary of Methodology Used by Brook et al. (1995a; 1995b) for RADM

1. Determination of categories

- Used Ward's (1963) method of cluster analysis, which minimizes within-cluster sums of squares, in an agglomerative, hierarchical mode for wind flow parameters. Used eastern North American zonal *u* and meridional *v* 85-kPa wind components for 0000 UTC with 5° spatial resolution (two variables at 48 grid nodes over three days).
- Clustered "consecutive" 3-day periods from 1979, 1981, and 1983, with subsequent classification of remaining "running" 3-day periods from 1979-1985 into one of these clusters.

2. Determination of number of clusters to retain

- Examined stepwise increases in within-group variance with decreasing number of clusters, expressed through F-statistic.
- Retained 19 clusters based on results for wind flow, sulfate deposition, and prior work by Fernau and Samson (1990a,b). Used quantitative information but somewhat subjective criteria.

3. Development of aggregation procedure

- Estimated total deposition for the group of sampled categories, from the sampled events (weighted sum, accounting for the number of events sampled from each cluster and frequency of occurrence of the clusters).
- Scaled up by ratio of total annual deposition to an estimate of annual deposition that is based on aggregating the sampled categories. Estimates used either mean deposition from a sampled category or deposition estimated from the deposition-precipitation relationship in the category.

4. Selection of "optimum" set of events

- Subdivided the 19 categories into wet and dry subsets, resulting in 38 new categories.
- Primarily represented categories that accounted for most of the annual wet sulfate deposition (at least 75% when combined). Selected 19 events from these categories. Also selected 11 events from winter (5) and dry (6) events to represent seasonal deposition differences, dry periods, and possible nonlinear effects.
- Number of events selected from each category was based on proportionality to the frequency of occurrence of the category and the percent of total sulfate accounted for by the category.
- Examined 20 potential sets (each randomly generated) meeting all criteria, selected the one that minimized RMSE for annual sulfate deposition (primarily) and precipitation (secondarily) at 13 sites. Selected sets in stages, first choosing a set of 19 events from the wet categories and then a set of 11 events from winter and dry periods.

17.1.2 Objectives

The present objectives differ somewhat from those which motivated the earlier research. RADM was primarily designed to address issues involving acidic deposition. CMAQ addresses a more diverse collection of air quality parameters with equal importance. In addition, CMAQ will be

applied to a continental domain that is significantly larger than geographic area for which estimates are supplied by RADM. The extension from the RADM domain to a continental domain is extremely ambitious as it relates to episode selection and aggregation. Therefore, the development of an approach that accommodates this larger continental domain is particularly challenging.

The objective of the activity described in this chapter is the development of an aggregation and episode selection approach that supports model-based annual and seasonal air quality estimates that are at least as accurate (with respect to sampling uncertainty) as those achieved by RADM, in consideration of the more general applicability envisioned for CMAQ both with respect to air quality parameters and geographic representation. This accuracy must be preserved while minimizing any additional cost. In essence, "cost" refers to the number of events for which the model provides estimates, each of which adds cost in the form of both computational processing time and human labor.

17.2 Summary of the Approach

The analysis was carried out in phases, with information gathered in each phase contributing to the design of the next. For this reason, this chapter is structured to present complete descriptions of the methodology and results achieved in each phase, in sequence, in sections 17.3, 17.4, and 17.5. This section is intended as an overview of the process prior to those detailed descriptions. This overview consists of descriptions of some key elements of the methodology, the rationale, scope, and limitations associated with the methodology, and the strategy used to move in phases toward the final result.

17.2.1 Basic Elements of the Methodology

Simply stated, the methods that we have employed involve the determination of meteorologically representative categories, the selection of events from those categories, and the use of evaluative tools to ensure that the detailed aspects of those activities are defined in such a way as to achieve optimal results, to the extent that we can measure optimality.

A specific goal is the definition of meteorological categories that account for a significant proportion of the variability exhibited by the air quality characterizations of interest. The basic approach used in the current analysis for the determination of categories and event selection components is related to that of Brook et al. (1995a,b), but certain fundamental considerations have been modified to reflect the differences inherent in the present objectives as described in section 17.1. The common element is the cluster analysis of zonal u and meridional v wind components to define meteorological categories.

The definition of meteorological categories is designed to support the selection of events from those categories in a process known as stratified sampling (Cochran, 1977). Stratified sampling exploits the internal homogeneity of the meteorological categories, or "strata," to achieve more precise estimates than would be possible using simple random sampling (i.e., randomly selecting

events without regard to meteorological category). Certain variations of stratified sampling are relevant to this analysis. One relatively inefficient option for invoking stratified sampling would involve selecting the same number of events from each category/stratum. This is known as "equal allocation." An alternative is "proportional allocation," which involves selecting numbers of events in direct proportion to the size of the stratum. Thus, more events are selected from strata that contain large numbers of events than from smaller strata. This is potentially much more efficient than equal allocation, in the sense that it leads to much more precise (i.e., lower variance) estimates. Estimates exhibiting absolute maximum efficiency (i.e., minimum variance) are obtained by modifying this method slightly so that the number of events selected from each stratum is in direct proportion to the product of the size of the stratum times the internal variability (as characterized by the standard deviation of the measurement of interest) within the stratum. Thus, strata exhibiting significant variability among events are sampled more heavily than strata in which events are more uniform. This is known as "optimum allocation," which is identical to proportional allocation when within-stratum variances are equal.

While wind flow parameters were used to define the meteorological categories, other meteorological parameters were used in subsequent phases of the analysis to refine aspects of the episode selection methodology, and as evaluative tools to assess the effectiveness of the approach. These parameters include visibility (as represented by the b_{ext}), temperature, and relative humidity. Their specific roles are discussed in more detail in the following sections.

17.2.2 Rationale, Scope, and Limitations

As stated previously, the approach to selecting an aggregation and episode scheme is based upon the definition of meteorological categories that account for a significant proportion of the variability exhibited by the air quality characterizations of interest. Strictly speaking these characterizations include parameters such as acidic deposition, air concentrations, and measures of visibility. Therefore, it might be argued that the definition of categories should be formulated directly using these parameters that are ultimately of interest. However, it is equally important that the model simulate the particular transport mechanisms involved in the associated atmospheric processes, and in particular that source-attribution analyses be facilitated. This requires that categories be defined with an emphasis on wind flow parameters. Indeed, characterizations of basic wind field patterns in essence describe frontal passages, along with all of the meteorological properties typically associated with them.

In view of the importance placed on the accurate simulation of transport mechanisms, a complete evaluation of the episode selection and aggregation methodology would require an assessment of the accuracy with which transport is represented. However, this accuracy cannot be measured, as there is no technique available to support a direct, quantifiable assessment of the representation of atmospheric transport. In addition, with the exception of b_{ext}, there are little air quality data available with the required spatial and temporal resolution and range to support a direct evaluation of the methods described in this chapter with regard to the outcome parameters that will be of primary interest in CMAQ.

For these reasons two meteorological parameters (in addition to b_{ext}), which are known to be related to many of the air quality parameters of interest, and for which appropriately resolved data are available, were used as evaluative measures in this analysis. Specifically, temperature and relative humidity were used; however, primary emphasis will be placed upon the b_{ext}, which provides a surrogate measure for fine particles. It must be recognized that this constitutes a secondary evaluative tool, in the sense that the effectiveness of the approach cannot be directly measured as it relates to atmospheric transport or to specific air quality parameters, both of which are primary outcomes. For this reason, the methods were not developed and refined solely toward the goal of optimizing performance associated with the estimation of visibility. Instead, this performance was evaluated in combination with other considerations that were believed to be important but for which performance may not be readily quantifiable.

17.2.3 Strategy

The basic strategy used for the selection of events to support aggregation-based estimation is described in the steps outlined below. The term "cluster" describes a collection of events that are defined to be meteorologically similar based upon cluster analysis results. The term "stratum" describes a collection of meteorologically similar events to be used in stratified sampling. In this chapter, "stratum" and "cluster" are essentially interchangeable since the clusters defined in the analysis are ultimately used as the strata for sampling purposes.

1. We explored different approaches to the cluster analysis of wind components, each of which resulted in the definition of a set of clusters (strata) of meteorologically similar events, for possible use in alternative stratification schemes. Some alternatives that were explored include annually defined clusters (i.e., strata defined using cluster analysis of daily wind field data from several years, without regard to season), seasonally defined clusters (i.e., strata defined using distinct cluster analyses of wind field data pertaining to different seasons), and regionally defined clusters (i.e., strata defined using distinct cluster analyses of wind field data for different geographic regions).

2. The alternative stratification schemes explored in step 1 were compared using relative efficiencies and meteorological considerations. The concept of relative efficiency relates to the variance associated with an estimate derived using different sampling schemes. We explored the variances associated with estimates of the annual means of the evaluative meteorological parameters described above (visibility, temperature, relative humidity). The relative efficiency of each stratification scheme is defined as the ratio of the variance associated with simple random sampling to the variance associated with stratified sampling using that scheme. A large relative efficiency is indicative of a high degree of precision (lower variance) associated with the estimate of interest. As discussed previously, since these evaluative meteorological parameters do not afford a complete, direct assessment of the validity of process with regard to the parameters to which it is ultimately to be applied, meteorological considerations were taken into account in combination with the relative efficiencies in order to select a general stratification

scheme. The selected scheme (seasonal stratification) was considered for more detailed refinement in step 3.

3. We determined an appropriate number of clusters to retain in combination with an acceptable number of events that would be necessary to achieve the objectives. The determination was based on estimated standard deviations associated with several alternative formulations, as well as other considerations. The standard deviations relate specifically to estimates of the annual means and 90th percentiles of the evaluative meteorological parameters. The "other considerations" included the objective of matching or exceeding the performance of the current aggregation methodology used for RADM, as well as a goal of avoiding unsampled clusters. Unsampled clusters (the inclusion of clusters for which no events would be selected) were considered to be undesirable because there would be no information available in the aggregation process to account for events contained in such a cluster. In concept, our view was that any unsampled cluster should be combined with another cluster to which it is most similar. In practice, we achieved this by constraining ourselves to collections of clusters that were adequate to support the sampling of at least one event from each cluster. Twenty clusters were ultimately retained, consisting of five clusters defined in each of the four seasons, and the number of events necessary to achieve the objectives was determined to be 40.

4. A stratified sample of 40 events was randomly selected from the 20 clusters defined in step 3. Proportional allocation was used in determining the number of events to be selected from each cluster (stratum). Optimum allocation was considered but was not used for several reasons, including: (1) it requires the quantification of the variance of a primary outcome parameter, whereas only secondary evaluative outcomes (visibility, temperature, relative humidity) were available as discussed previously, and (2) the variance of any outcome parameter varies geographically, so that optimum allocation would likely result in differing numbers of events depending upon the geographic location, whereas proportional allocation would not. It was verified that, at least based upon the evaluative outcome parameters that are available, in most geographic locations the distribution of events that arises from proportional allocation does not differ substantially from that arising from optimum allocation.

Details concerning the implementation of this approach are discussed in the following sections. Sections 17.3 and 17.4 correspond to steps 1 and 2 in the strategy outlined above, and section 17.5 includes a discussion of steps 3 and 4.

17.3 Cluster Analysis of Wind Fields

The cluster analysis of wind components is described in this section. This includes a description of the wind field data, the basic analysis technique, variations of the basic technique, and methods of presentation. Some graphical results of exploratory analyses are also included.

17.3.1 Description of Wind Data

To accommodate the continental domain and to achieve adequate spatial resolution, the cluster analysis involves data at 336 grid nodes with 2.5° spatial resolution, as obtained from the NCEP/NCAR 40-year reanalysis project (Kalnay et al., 1996). In this analysis, 700 mb wind components for 1800 UTC have been used, in consideration of the mountainous western regions in the domain. Corners of the grid were cut back to guard against excessive influence from ocean-based meteorology. Graphical illustrations of this domain are referenced later in this section.

17.3.2 Basic Cluster Analysis Technique

Cluster analysis, in the present formulation, involves the classification of a set of observations into categories that are internally homogeneous with respect to defined multivariate relationships in the data. In this case "multivariate" refers to the multiple variables used to characterize wind fields, consisting of the *u* and *v* components applicable to each location within the previously described domain and extending over an event that includes multiple days.

A 12-year period (1984-1995) was considered in the exploratory cluster analyses, later refined to a 9-year period (1984-1992) for the final clustering upon which episode selection was based. Since CMAQ is actually run for a 5-day period for each event (the first two days establish initial conditions, and model predictions from days 3-5 are saved as a "3-day event"), 5-day periods were clustered rather than 3-day periods. To make the analysis computationally feasible, the first, third, and fifth days of each 5-day event were considered. Based upon the performance noted by Fernau and Samson (1990a,b), Ward's method of cluster analysis was used (Ward, 1963), minimizing within-cluster sums of squares, in an agglomerative (i.e., moving from many clusters toward fewer clusters), hierarchical (i.e., once clusters are joined they cannot be separated) process. Thus, if a single observation (event) is considered to consist of 2,016 elements (the 2 *u* and *v* components × 336 grid nodes × 3 days considered per event), then the objective of the cluster analysis is to divide these observations into clusters (categories) for which the within-cluster sum of squares (sum of squared differences between the elements of individual observations or means) is minimized.

In the exploratory analyses, clusters were initially defined based upon "consecutive" rather than "running," or overlapping 5-day periods from 1987-1992. Then, each remaining event ("running" 5-day periods from 1984 through 1995) 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* component from the cluster mean *u* and *v*. In the final cluster analysis, using 1984-1992 data, consecutive 5-day periods from 1984 through 1992 were clustered, and remaining events were classified into those clusters according to the same criteria described above. Cluster analyses were carried out using SAS (SAS Institute Inc., 1989); however, due to the extreme computational burden of these analyses it was necessary to calculate the distance matrix externally from the clustering procedure itself.

17.3.3 Illustration of Cluster Analysis Results

Representative results from the selected exploratory analyses, as well as all results from the final cluster formulation, are illustrated graphically. Specifically, cluster definitions are illustrated using maps of cluster average wind fields. In some cases, these are supplemented with specialized maps illustrating the intracluster variability in the wind fields. Star chart histograms are also used to illustrate the frequency of occurrence of events from each cluster, for each month of the year.

To provide proper perspective for a review of this analysis, it is useful to consider preliminary results for a set of 30 clusters initially defined using annual data from 1984-1995. The most prevalent cluster (labeled Cluster 1) accounted for 12.19% of all 5-day events during this time period, with nearly all of those events occurring during the summer months. Map-based graphs of mean wind vectors for this cluster are contained in Figure 17-1 (a-c). Three maps are included, representing mean vectors for days 1, 3, and 5 of the 5-day event. These maps illustrate largely stagnant conditions associated with an anticyclonic pattern centered over the southeastern U.S.; a zonal flow over southern Canada and a trough over the west coast of the U.S.

As a second example, Figure 17-2(a-c) illustrates mean wind vectors for a somewhat less prevalent cluster that accounted for 3.65% of all 5-day events during this time period (ranking ninth in overall prevalence and labeled as Cluster 9). Most of these events occurred between the months of October and April, characterized by northwesterly winds associated with a large-scale trough moving through the central and eastern portion of the domain.

While these maps are effective in illustrating average behavior associated with each cluster, they do not give any indication of the variability inherent in the clusters. Figures 17-4 (a-c) and 17-5 (a-c) contain additional maps that address this issue. The first map (Figure 17-3a) illustrates the wind field for the third day of an individual event (January 23-27, 1989) that was assigned to Cluster 9. Comparison of this map to the mean wind field for day 3 of Cluster 9 (Figure 17-2b) reveals fairly close resemblance between the two. By contrast, the maps in Figures 17-3b and 17-4c depict the third day from two other events belonging to Cluster 9. These patterns do not resemble the mean wind field as closely, and are indicative of the variability among individual events that were assigned to this cluster.

Clearly it would be useful to simultaneously visualize the variability exhibited by all of the wind fields assigned to each day of Cluster 9, and Figure 17-4 (a-c) represents an attempt to do this. These maps contain the mean wind vectors for each day of Cluster 9 on a thinned-out grid that only includes alternating grid nodes. In addition, the maps contain groups of small dots, each of which depicts the location of the wind vector arrowhead for an individual event assigned to this cluster. The groups of dots collectively illustrate the distribution of arrowheads for all events belonging to Cluster 9.

The dots surrounding each mean wind vector arrowhead appear in a somewhat circular pattern, and the extent of spread exhibited by the dots illustrates the degree of variability among wind vectors assigned to the cluster. Similar patterns characterize the variability associated with other clusters (not shown). Clearly there is substantial variability associated with the wind vectors assigned to individual clusters. This variability emphasizes the ambitious nature of the endeavor. In essence, the goal is to categorize many years worth of meteorological patterns into a finite number of classes. Furthermore, each meteorological pattern does not simply describe a single location at a given point in time; it is required to simultaneously represent a broad spatial domain over a significant temporal period. Indeed, it should not be surprising that a substantial amount of variability is associated with the result.

Figure 17-5 (a-e) contains star chart histograms of the 30 clusters defined using annual data. Each chart illustrates the frequency of 5-day events belonging to a given cluster, and the clusters themselves are ordered according to overall frequency of occurrence. As shown on the charts, events from Cluster 1 accounted for 12.19% of all 5-day events between 1984 and 1995, those from Cluster 2 accounted for 12.10% of all events, etc. 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 based upon these charts:

- Although defined using annual data, the cluster frequencies reveal definite seasonal tendencies. That is, clusters do not occur randomly throughout the year, but rather exhibit a tendency to occur more frequently within specific seasons. Thus, the annual 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 both spring and fall months.

- The two most prevalent clusters heavily emphasize the summer months; each of these clusters includes more than twice as many events as any other cluster, with the vast majority of summer events contained in them.

The disproportionate representation of summer events by two of the 30 clusters is not surprising, since the wind fields are expected to be less variable in the summer. However, seasonal differences in meteorology and atmospheric chemistry are important in explaining the variability exhibited by the air quality parameters of interest. The addition of more warm season clusters could provide improved resolution in this regard. This is the primary motivation for conducting analyses using seasonally distinct clustering.

17.4 Evaluation of Alternative Aggregation Approaches

The previous illustration clarifies the motivation for investigating seasonally distinct clustering. Similarly, regionally distinct clustering offers an alternative that might provide improved resolution of wind field groupings on smaller spatial scales. An added dimension to the problem is that the number of clusters to be retained clearly affects the degree of resolution. To gain an understanding of the importance of these considerations as they relate to estimation of the meteorological parameters used as evaluative tools in this analysis, these and other alternatives were explored in several combinations. These explorations, designed to expose patterns and trends from which to make an informed selection of a general approach, are described in this section.

17.4.1 Description of Alternative Approaches

Several variations of the *u* and *v* wind vector clustering were investigated. As discussed previously, these were selected to investigate patterns and properties, and not necessarily to be considered as final candidates for cluster definitions. They are as follows:

- Annually defined strata with variations in the number of strata: 5, 10, 20, 30, 60, and 90 clusters;

- Seasonally defined strata using warm and cold seasons: totals of 10, 20, 30, and 60 clusters equally divided between the warm season April-September period and the cold season October-March period;

- Seasonally defined strata approximately equally divided among summer (June, July, August), winter (December, January, February), and transitional (spring and autumn combined) seasons: totals of 20, 30, and 60 clusters;

- Seasonally defined strata equally divided between summer and winter seasons, and with approximately twice as many strata defined for the transitional season: totals of 20 and 30 clusters;

- Seasonally defined strata approximately equally divided among summer, winter, spring (March, April, May), and autumn (September, October, November) seasons: totals of 20 and 30 clusters; and

- Regionally defined strata: 30 clusters identified in each of four separate cluster analyses for the northeast, southeast, northwest, and southwest subsets of the domain

We also investigated the alternatives of strata defined to simply mirror the four seasons (disregarding all clusters), strata defined by clustering of 3-day events rather than 5-day events, and strata defined by clustering the meteorological parameters used to evaluate the approach.

17.4.2 Description of Meteorological Data

For the alternative stratification schemes, preliminary testing was performed by examining the uncertainty associated with the use of cluster-based stratified sampling to estimate the annual mean of daily noontime levels of visibility, temperature, and relative humidity, with primary emphasis placed upon visibility as discussed previously. Visibility (units of km^{-1}) was specifically expressed as the light extinction (b_{ext}), less observations with precipitation, and less observations with relative humidity greater than 90%. The light-extinction coefficient is often used to characterize visibility, although in general, it has limited ability to predict human visibility. The visual range v_r (km) can be estimated from the b_{ext} by using the Koschmieder equation:

$$v_r = \frac{3.91}{b_{ext}} \qquad (17\text{-}1)$$

Temperature is in units of degrees Celsius, and relative humidity in percent. These parameters were taken from 201 locations in the continental U.S. for which coverage was at least 99%, as illustrated in Figure 17-6. This was specifically defined as sites exhibiting at least 99% completeness for light extinction coefficient; allowances were made for missing observations that were associated with precipitation so as not to bias the inclusion of sites toward drier climates.

17.4.3 Analysis Methods

The average relative efficiency associated with the estimation of mean annual visibility, temperature, and relative humidity, using each alternative stratification scheme, was used for comparison of the schemes. Specifically, at each location and for each scheme, the variance of an aggregation-based estimate (Cochran, 1977) of the annual mean was determined in three ways, assuming that the estimated mean was calculated using (1) a stratified sample with equal allocation across strata, (2) a stratified sample with proportional allocation across strata, and (3) simple random sampling, with a total sample size consisting of the same numbers of events in each case. Also at each location, the ratio of the simple random sampling variance to the variance associated with each of the two stratified sampling designs was calculated and expressed as the relative efficiency of each of those designs. Finally, those relative efficiencies were averaged across sites to provide an indication of the overall efficiency of each scheme.

This is best explained using a specific example. First, suppose that the mean annual temperature at a given location is estimated as the average of the daily temperatures from 30 randomly sampled 3-day events from the period 1984-1992, completely disregarding any information related to clusters. Suppose that the standard deviation associated with that estimate is 1.5°C, so that a 95% confidence interval would yield the estimated mean ±2.94°C (=1.96×1.5). Second, suppose that the mean is instead estimated as a weighted average from 30 3-day events, using two events per cluster (i.e., stratified sampling with equal allocation), and using the frequency of occurrence of each cluster as the weight applied to the temperature from the corresponding event. Suppose that the standard deviation associated with that estimate is 1.0°C, compared to

1.5°C from simple random sampling. Last, suppose that 30 events are selected using proportional allocation (i.e., the number of events selected from a cluster is proportional to the number of events belonging to the cluster), and that the standard deviation of the resulting estimated annual mean temperature is 0.8°C. These standard deviations (1.5, 1.0, and 0.8) translate to variances of 2.25, 1.0, and 0.64 for simple random sampling, stratified sampling with equal allocation, and stratified sampling with proportional allocation, respectively. Thus, for this hypothetical location, the relative efficiency of stratified sampling with equal allocation is 2.25/1.0=2.25, and the relative efficiency of stratified sampling with proportional allocation is 2.25/0.64=3.52. Proportional allocation is more efficient than equal allocation in the sense that it leads to lower variances and therefore tighter confidence intervals bounding the estimated mean.

17.4.4 Results

Tables 17-2 (a-b) and 17-3 present mean relative efficiencies associated with annual means of the daily noontime temperature, relative humidity, and extinction coefficient, as estimated using aggregation approaches based upon the various schemes described above. In each table, results are presented to illustrate the relative efficiency of estimation methods using equal allocation (equal numbers of events selected to represent each cluster) and proportional allocation (numbers of events selected in direct proportion to the total number of events categorized into the given cluster). Relative efficiencies reported in this section are valid for any number of events that might be selected, as relative efficiency is invariant to sample size under equal or proportional allocation

In the case of proportional allocation, the relative efficiency actually refers to the minimum variance that might theoretically be achieved if proportional allocation were carried out precisely. In practice this might only be possible if a very large number of events were to be sampled, since the appropriate proportions might otherwise dictate sampling of fractional numbers of events from some clusters. Therefore, the relative efficiency reported for proportional allocation may be thought of as an upper limit to the relative efficiency that might actually be achieved in practice. In all likelihood, this upper limit cannot be attained, but it should be possible to achieve a relative efficiency occurring somewhere within the range defined by this upper limit and the relative efficiency associated with equal allocation of events among clusters.

The first six rows in Table 17-2a illustrate relative efficiencies associated with various numbers of annually defined strata (i.e., clusters emerging from cluster analyses of daily wind field data from 1984-1995 without regard to season). Several observations may be made by inspecting the first six rows in Table 17-2a:

- The relative efficiency associated with the estimation of mean temperature is consistently highest, followed by that of relative humidity. This implies that the meteorological clusters are most useful in distinguishing between events with regard to temperature, and least useful in distinguishing extinction coefficient.

- In the case of temperature, the variability associated with stratified sampling using wind field-based clusters is consistently less than the variability associated with simple random sampling (i.e., relative efficiencies are uniformly greater than 1.0). Thus, in each scheme the clusters contribute important information that explains variation in temperature.

- For the estimation of mean relative humidity and mean extinction coefficient, the use of stratified sampling based upon wind field clusters is consistently more efficient than simple random sampling if proportional allocation is approximately satisfied. In most cases, the use of equal allocation actually results in less efficient estimation than simple random sampling; this reflects a wide range of stratum sizes that would be inappropriately represented using equal allocation.

- As would be expected, the efficiency associated with proportional allocation increases as the number of strata increases, incorporating more refined representations of the meteorological classes within the strata.

- The efficiency associated with equal allocation decreases as the number of strata increases. The stratum sizes become more divergent when more strata are defined, so that the inefficiency of equal allocation is magnified.

Based upon the well-known properties of stratified sampling discussed above, our objective was to design and implement a scheme based upon approximate proportional allocation. The numbers for equal allocation in these tables merely served to provide a lower bound on the efficiency that could be realized, since it was known that the precise degree of efficiency reported for proportional allocation might not be achievable in practice due to the requirement of sampling integer-valued numbers of events.

As discussed in section 17.3.3 above in regard to the analysis illustrated in the "30 Strata" row of the table, summer events are disproportionately represented by two of the 30 clusters. To evaluate the effect of improving the resolution of summer meteorological patterns, as well as the effects of imposing constraints that would alter the resolution of families of clusters under a variety of scenarios, various implementations of seasonally distinct clustering were investigated. The last four rows of Table 17-2a illustrate results associated with stratum definitions based upon a simple warm/cold seasonal dichotomy, in which separate cluster analyses were conducted to force equal numbers of strata for each season. The following observations may be made based upon this portion of the table:

- With the exception of extinction coefficient estimation under proportional allocation, stratification using seasonally defined clusters consistently yields improved efficiency over stratification using the same number of annually defined clusters. Thus, although the annually defined clusters do adhere to a seasonal pattern, the improved resolution afforded by the forced inclusion of more warm weather clusters (and reduction of cold weather clusters) is particularly effective in explaining variation in temperature.

- For seasonally defined clusters, the relative gains in efficiency associated with using equal allocation are large as compared to the potential gains associated with using proportional allocation. Thus, in departing from proportional allocation (which is not precisely achievable in practice as discussed above), seasonally defined clusters are likely to afford improved efficiency over annually defined clusters.

Results associated with further seasonal stratification schemes are illustrated in Table 17-2b. The first two rows correspond to approximately equal numbers of clusters divided among summer, winter, and transitional (spring and autumn combined) seasons, the next two rows correspond to equal numbers of clusters divided between summer and winter with approximately twice as many transitional season clusters, and the fifth and sixth rows correspond to approximately equal numbers of clusters divided among four seasons. In each case, the exact distributions are constrained to result in total numbers of strata that are directly comparable to the numbers of strata investigated in other seasonal and annual analyses.

Under proportional allocation, stratification schemes based on three or four seasons offer significantly improved efficiency in the estimation of mean temperature compared to two-season and annual stratification schemes with comparable total numbers of strata. They also demonstrate slight but uniform improvement in the estimation of extinction coefficient, and mixed results in the estimation of relative humidity.

Table 17-2a. Mean relative efficiency[1] associated with estimation of the annual (1984-1995) mean of the indicated parameter, using various stratified[2] sampling[3] approaches relative to simple random sampling. *Continental analysis.*

	Temperature		Relative Humidity		RH-Adjusted Extinction Coefficient	
Method	**Equal Allocation Across Strata**	**Proportional Allocation**	**Equal Allocation Across Strata**	**Proportional Allocation**	**Equal Allocation Across Strata**	**Proportional Allocation**
5 Strata	2.34	2.41	1.14	1.18	1.01	1.12
10 Strata	2.29	2.50	1.09	1.22	0.90	1.14
20 Strata	2.11	2.83	0.88	1.26	0.69	1.16
30 Strata	2.08	2.86	0.86	1.28	0.67	1.17
60 Strata	1.81	3.05	0.76	1.32	0.59	1.19
90 Strata	1.43	3.10	0.62	1.33	0.53	1.20
10 Strata Defined Seasonally (5 Warm, 5 Cold)	2.62	2.94	1.11	1.26	0.93	1.13
20 Strata Defined Seasonally (10 Warm, 10 Cold)	2.91	3.54	1.11	1.35	0.89	1.16
30 Strata Defined Seasonally (15 Warm, 15 Cold)	3.14	3.60	1.17	1.37	0.91	1.17
60 Strata Defined Seasonally (30 Warm, 30 Cold)	2.74	3.89	0.98	1.39	0.77	1.20

[1] Relative efficiency is defined as the ratio of the variance associated with simple random sampling to the variance associated with stratified sampling. The table entries are means of station-specific efficiency ratios, averaged across stations within the continental domain.

[2] Unless otherwise noted, stratum definitions are based on annual clustering of 5-day events from a temporal subsample of the wind field data; the remainder of the sample is then classified into those strata.

[3] Reflects sampling of 3-day events from 1984-1995.

Table 17-2b. Mean relative efficiency[1] associated with estimation of the annual (1984-1995) mean of the indicated parameter, using various stratified sampling[2] approaches relative to simple random sampling. *Continental analysis.*

Method	Temperature		Relative Humidity		RH-Adjusted Extinction Coefficient	
	Equal Allocation Across Strata	Proportional Allocation	Equal Allocation Across Strata	Proportional Allocation	Equal Allocation Across Strata	Proportional Allocation
20 Strata Defined Seasonally (6 Summer, 6 Winter, 7 Transitional)	2.66	3.91	0.95	1.31	0.90	1.18
30 Strata Defined Seasonally (10 Summer, 10 Winter, 10 Transitional)	2.60	4.12	0.87	1.35	0.78	1.20
20 Strata Defined Seasonally (5 Summer, 5 Winter, 10 Transitional)	3.31	4.06	1.06	1.33	0.89	1.19
30 Strata Defined Seasonally (8 Summer, 8 Winter, 14 Transitional)	2.91	4.17	0.96	1.36	0.80	1.20
20 Strata Defined Seasonally (5 Summer, 5 Winter, 5 Spring, 5 Fall)	3.23	3.86	1.11	1.36	0.92	1.18
30 Strata Defined Seasonally (8 Summer, 8 Winter, 7 Spring, 7 Fall)	2.82	3.88	1.03	1.38	0.84	1.19
4 Strata Defined as Seasons (Dec-Feb, etc.)	2.93	2.93	1.21	1.21	1.12	1.12
30 Strata Defined by Clustering 3-Day Events	2.24	2.78	0.97	1.29	0.76	1.16
30 Strata; Seasonality Removed from Met. Parameters	0.92	1.28	0.74	1.15	0.63	1.09
30 Strata Defined by Clustering Lt. Ext. Coeff.	0.51	1.89	0.30	1.11	0.57	1.45

[1] Relative efficiency is defined as the ratio of the variance associated with simple random sampling to the variance associated with stratified sampling. The table entries are means of station-specific efficiency ratios, averaged across stations within the continental domain.

[2] Reflects sampling of 3-day events from 1984-1995.

The final four rows of Table 17-2b address issues that are useful in that they provide additional perspective for this analysis. They are discussed in sequence below:

- In view of the improvement associated with seasonal stratification schemes relative to annual stratification, a natural question to ask is whether a seasonal scheme with one stratum per season (i.e., no cluster analysis) is sufficient to support comparable efficiency. Based upon the results in Table 17-2b, this method offers overall improvement relative to an annual scheme with an approximately equal number of strata (i.e., five strata as shown in the first row of Table 17-2a). As might be expected, it is significantly less efficient than the use of 20 strata equally divided among four seasons, but not dramatically so.

- Another natural question is whether the clustering of five-day events (using wind data from the first, third, and fifth days) has a noticeable impact on this analysis relative to the clustering of three-day events that previous investigators have pursued. Based upon results in Table 17-2b, the definition of 30 strata using annual clustering of three-day events produces similar results to those associated with 30 strata using five-day event clustering (fourth row of 17-2a) with respect to the evaluative parameters. Note that this does not necessarily address differences with respect to the characterization of transport.

- A consistent result in these analyses is that the relative efficiency associated with the estimation of mean temperature is much greater than that associated with the estimation of mean relative humidity, which in turn is slightly greater than the relative efficiency associated with extinction coefficient. Since the preliminary results demonstrate associations between the clusters and the seasons (even for annually defined clusters), one might hypothesize that this characteristic is related to the more pronounced seasonal trends associated with temperature and the less pronounced trends associated with extinction coefficient. Table 17-2b illustrates relative efficiencies associated with the three meteorological parameters following the removal of seasonal trends from each. (Trends were removed by analyzing deviations of each parameter from a sinusoidal curve fitted to the raw data at each location.) Indeed, the relative efficiencies for the three parameters are much more comparable in this context, and much more similar to the results for extinction coefficient in other analyses. This analysis lends support to the use of extinction coefficient as the primary evaluative outcome, because it reflects the ability of each scheme to characterize short-term meteorological patterns apart from long-term seasonal trends.

- A final investigation in Table 17-2b also relates to the utility of extinction coefficient as the primary evaluative outcome. Under proportional allocation, the relative efficiencies associated with it are not dramatically greater than 1.0, indicating that stratified sampling based upon wind field clustering produces consistent but not dramatic gains in efficiency (relative to simple random sampling) in the estimation of mean extinction coefficient. To put this observation in its proper perspective, it is useful to consider the maximum relative efficiency that might be achieved from any stratified analysis. In particular, a

stratification of events was performed based upon extinction coefficient itself (rather than wind fields). Under these optimal conditions, the relative efficiency associated with a 30-stratum scheme was still only 1.45, compared to a range of 1.17–1.20 for other 30-stratum schemes based upon wind field clusters. Considered in this context, relative efficiencies encountered in these tables are encouraging. (This clustering of extinction coefficient provides some useful perspective regarding this evaluative outcome parameter, but is not pursued as a recommended methodology for estimation based upon the rationale outlined in section 2.2.)

To evaluate the utility of regionally distinct stratification schemes, the continental domain was divided into four quadrants and cluster analyses were performed on wind fields within each region, resulting in four sets of 30 annually defined strata. For each region, relative efficiencies were summarized as in the previously described analyses. Since the mean relative efficiencies are constrained to sites within each region, results in Table 17-2a are inappropriate for comparison to these regional mean relative efficiencies. Therefore, Table 17-3 also includes mean relative efficiencies only for the sites within each region, from comparable clustering of continental data. These are displayed in combination with the results for strata defined using regional data.

For example, using 30 strata under proportional allocation, clustering of wind fields in the northeast quadrant of the domain results in a mean relative efficiency (averaged over sites in that quadrant) of 1.65 associated with the estimation of mean temperature. For 30 strata, from clusters defined over the entire continental domain, the mean relative efficiency, over those same northeast sites is 3.14. The results in Table 17-3 collectively demonstrate that, under proportional allocation, regional stratification produces either no gains or only slight gains in efficiency in both the southeast and southwest regions for any of the evaluative parameters. In the northeast and northwest regions, this technique is markedly less efficient than continental clustering with regard to temperature, and somewhat less efficient with regard to relative humidity. There is only a negligible effect on extinction coefficient.

This result has significant importance, because it demonstrates that in the northern half of the domain, clustering of continental wind field data is actually more effective than clustering regional data in explaining variation in some meteorological parameters on a regional scale. Thus, the wind field patterns associated with different levels of temperature (and, less markedly, relative humidity) in a northern region are more distinctly identified on a continental scale than on a regional scale. Furthermore, wind field patterns associated with different levels of extinction coefficient (a primary evaluative parameter due to its close association with fine particles) in a given region are no more distinctly identified on a regional scale than on a continental scale.

Table 17-3. Mean relative efficiency[1] associated with estimation of the annual (1984-1995) mean of the indicated parameter, using various stratified sampling[2] approaches relative to simple random sampling. *Regional analysis.*

	Temperature		Relative Humidity		RH-Adjusted Extinction Coefficient	
Region/Method	Equal Allocation Across Strata	Proportional Allocation	Equal Allocation Across Strata	Proportional Allocation	Equal Allocation Across Strata	Proportional Allocation
Northeast; 30 Strata Def. Using Regional Data	1.41	1.65	0.98	1.15	0.95	1.27
Northeast; 30 Strata Def. Using Continental Data	2.27	3.14	0.77	1.20	0.61	1.28
Southeast; 30 Strata Def. Using Regional Data	2.35	2.72	1.04	1.32	0.77	1.22
Southeast; 30 Strata Def. Using Continental Data	2.02	2.62	0.83	1.24	0.61	1.20
Southwest; 30 Strata Def. Using Regional Data	1.95	2.42	0.94	1.26	0.81	1.01
Southwest; 30 Strata Def. Using Continental Data	1.88	2.43	0.89	1.22	0.84	1.04
Northwest; 30 Strata Def. Using Regional Data	1.17	1.66	0.92	1.27	0.82	1.08
Northwest; 30 Strata Def. Using Continental Data	1.96	2.82	1.01	1.47	0.71	1.03

[1] Relative efficiency is defined as the ratio of the variance associated with simple random sampling to the variance associated with stratified sampling. The table entries are means of station-specific efficiency ratios, averaged across stations within the continental domain.

[2] Reflects sampling of 3-day events from 1984-1995.

In view of the current objective of matching or exceeding the performance of the current RADM stratification scheme over the continental domain, it is appropriate to compare the relative efficiency of the stratification scheme used in the aggregation of RADM output compared to the relative efficiency of the schemes defined above. Table 17-4 addresses this specific issue. In addition to the results associated with equal and proportional allocation, this table illustrates the mean relative efficiencies associated with the actual allocation of the 30 RADM events, which lies somewhere between those extremes. (Proportional allocation was not a specific goal of the methods developed for RADM, although it is not inconsistent with the less formally stated goal of selecting more events from the clusters that accounted for most of the acidic deposition.)

Table 17-4. Mean relative efficiency[1] associated with estimation of the annual (1982-1985) mean of the indicated parameter, using various stratified sampling[2] approaches relative to simple random sampling. *RADM comparative analysis.*

Method	Temperature			Relative Humidity			RH-Adjusted Extinction Coefficient		
	Equal Alloc.	Proportional Alloc.	Actual Alloc. of RADM Events	Equal Alloc.	Proportional Alloc.	Actual Alloc. of RADM Events	Equal Alloc.	Proportional Alloc.	Actual Alloc. of RADM Events
19 Existing RADM Strata Defined from Clustering 1979-83 3-Day Events	0.98	1.58	1.45	0.76	1.18	1.07	0.67	1.18	1.11
38 Existing RADM Strata Defined After Separating Wet & Dry Categories	0.83	1.76	1.12	0.59	1.19	0.88	0.49	1.17	0.92
4 Strata Defined as Seasons (Dec-Feb, etc.)	3.06	3.05	2.77	1.13	1.13	1.02	1.17	1.18	1.11

[1] The table entries are means of station-specific efficiency ratios, averaged across stations within the RADM geographic domain.

[2] Reflects sampling of 3-day events from 1982-1985.

In the RADM development, 19 strata were defined based upon clustering of 3-day events occurring between 1979 and 1983. These were then subdivided into wet and dry categories, ultimately resulting in twice as many (38) strata that are actually employed in aggregation-based estimation. (Events from the 1982-1985 time period were classified into these strata, with 30 events selected for use in RADM.) The temporal and spatial domains applicable to the RADM development differ from those of the current analysis; therefore, any comparisons should be considered in this context.

The RADM domain approximately corresponds to the combined northeast and southeast regions displayed in Table 17-3. A comparison of Table 17-4 to Table 17-3 results under proportional allocation suggests superior performance associated with the strata identified using clustering of continental wind field data, with respect to the outcome measures used here. When the actual allocation of RADM events is considered, the superiority of proportional allocation associated with the continental analysis is further elevated.

17.5 Refinement of the Sampling Approach

Based, in part, on the results discussed in section 17.4, a stratified sampling scheme involving seasonal clustering based upon four distinct seasons was selected for further consideration and refinement. The general superiority of three- and four-season stratification schemes was discussed previously in relation to results depicted in Table 17-2 (a-b), and a four-season scheme was selected following additional considerations regarding differences in emission patterns

between spring and autumn that would not be apparent using our evaluative parameters alone. Another benefit of using this type of scheme is that it naturally lends itself to the development of seasonal estimates based upon four-season partitions. The derivation of such estimates from a two- or three-season scheme would not be as well defined, and the estimates themselves would likely be less precise.

Having selected this general approach, refinements were needed to determine an appropriate number of strata, and to arrive at an adequate number of events for sampling. These refinements, and a description of the sample of events that ultimately was selected, are discussed in this section. This aspect of the analysis was limited to a refined time frame consisting of the nine-year period from 1984-1992, which was targeted to ultimately represent baseline meteorology for use in modeling.

17.5.1 Determination of Appropriate Numbers of Strata and Events

The analysis proceeded under the assumption that, in order to satisfy the goal of matching or exceeding the performance of the RADM 30-event stratification scheme, a minimum of 30 events would be required in the framework of the continental domain. The precision associated with the estimation of annual means of the evaluative parameters was investigated for a range of 30 to 60 events, and for 16, 20, 24, and 28 seasonally defined strata (4, 5, 6, and 7 strata per season, respectively).

These numbers of strata were chosen as candidates because 19 wind field-based clusters were defined in the RADM scheme, which offers a certain degree of resolution with regard to the characterization of transport (which is not specifically addressed by the evaluative parameters investigated here). The range of 16 to 28 strata was selected to provide comparable resolution of wind fields and associated transport, noting that higher numbers of strata result in greater variation in the sizes of those strata, and this would force a more pronounced deviation from the goal of proportional allocation. As discussed previously, a primary goal was to ensure that every stratum is sampled, i.e., that there are no clusters which go unrepresented in the final set of events.

Note that we did not adhere to traditional rules of thumb regarding the determination of appropriate numbers of clusters to retain. These rules are based upon an assumption that some finite number of clusters is appropriate to represent the variability inherent in these patterns, and that additional clusters beyond that point add relatively little information. In reality, clusters defined during this process represent a continuum, and traditional F-test statistics illustrate this continuum quite smoothly. There is no magic number of clusters after which the relative importance of additional clusters drops noticeably. Indeed, as cluster analysis is used here merely for the definition of strata and not as an end in itself, there is no compelling reason to be restricted by existing conventions regarding determinations of an optimal number of clusters.

Standard deviations associated with estimation under all of the combinations described above are illustrated in Table 17-5. Furthermore, the process was repeated after seasonal adjustment of

each outcome parameter (seasonality was removed by analyzing deviations from a fitted sinusoidal curve), to ensure that long-term seasonal trends do not unduly influence any pattern. The standard deviations in Table 17-5 provide an indication of the degree of precision associated with estimation, and are influenced by the observed within-stratum variability of the parameter. One would expect the inclusion of more strata to facilitate greater precision, unless this increased precision is offset by incurred deviations from proportional allocation.

Table 17-5. Standard deviation[1] associated with estimation of the annual mean of the indicated parameter, using stratified sampling[2] with 16, 20, 24, or 28 strata and 30, 35, 40, 45, 50, 55, or 60 events. *Continental analysis.*

No. of Events	Temperature, deg. C				Relative Humidity, %				RH-Adjusted Extinction Coefficient, km^{-1} ($\times 10^3$)			
	16 Strata	20 Strata	24 Strata	28 Strata	16 Strata	20 Strata	24 Strata	28 Strata	16 Strata	20 Strata	24 Strata	28 Strata
	NOT SEASONALLY ADJUSTED											
30	0.97	0.98	0.95	0.95	2.38	2.40	2.38	2.40	6.69	6.73	6.73	6.77
35	0.89	0.89	0.87	0.87	2.19	2.20	2.20	2.21	6.03	6.08	6.10	6.23
40	0.83	0.83	0.81	0.81	2.04	2.05	2.06	2.05	5.71	5.63	5.70	5.71
45	0.79	0.77	0.76	0.75	1.93	1.92	1.93	1.93	5.38	5.34	5.33	5.36
50	0.74	0.74	0.72	0.72	1.82	1.83	1.82	1.82	5.09	5.06	5.07	5.08
55	0.71	0.70	0.69	0.68	1.74	1.74	1.73	1.73	4.78	4.79	4.83	4.83
60	0.67	0.67	0.66	0.65	1.66	1.66	1.65	1.65	4.62	4.62	4.60	4.60
	SEASONALLY ADJUSTED											
30	0.75	0.76	0.75	0.75	2.32	2.34	2.33	2.35	6.67	6.70	6.70	6.75
35	0.69	0.69	0.69	0.69	2.14	2.15	2.16	2.16	6.01	6.05	6.08	6.21
40	0.64	0.64	0.64	0.64	1.99	2.01	2.02	2.01	5.69	5.61	5.68	5.69
45	0.61	0.60	0.60	0.60	1.89	1.88	1.89	1.88	5.36	5.32	5.31	5.34
50	0.57	0.57	0.57	0.57	1.78	1.79	1.78	1.78	5.07	5.04	5.05	5.06
55	0.54	0.54	0.54	0.54	1.70	1.70	1.70	1.69	4.77	4.78	4.81	4.81
60	0.52	0.52	0.51	0.51	1.62	1.62	1.62	1.62	4.60	4.60	4.59	4.59

[1] The table entries reflect the standard deviation associated with the average station-specific variances (averaged across stations within the continental domain).

[2] Results reflect sampling of 3-day events from 1984-1992.

For temperature and relative humidity, the table suggests that for any given number of events, any effect associated with different numbers of strata is negligible. For extinction coefficient,

standard deviations for 30 and 35 event schemes increase slightly with increasing numbers of strata. There is a minimum standard deviation associated with 20 strata for the 40-event scheme, and only negligible effects across strata for greater numbers of events.

The objective of this determination was to develop a sufficiently large set of strata to provide some resolution with regard to the characterization of transport, yet be sufficiently concise to avoid any reduction in precision that would result from unsampled strata or from the inability to approximately satisfy proportional allocation. In consideration of these criteria and the above results, 20 strata were determined to constitute an appropriately sized set.

Figure 17-7 (a-c) provides perspective regarding the relative precision provided by sample sizes of 25 or more events associated with estimation of annual means of the evaluative parameters, based upon the use of 20 seasonally defined clusters as strata. The standard deviation displayed in each graph is actually associated with the average of the variances across sites. These graphs illustrate the advantage of stratified sampling with proportional allocation relative to simple random sampling. They also illustrate the relative gains in precision (expressed as reduction in standard deviation) that are realized as the number of events is increased. The star symbols plotted on the graphs indicate the actual standard deviation that would be realized for selected sample sizes under a 20-stratum scheme, with events distributed in accordance with proportional allocation to the extent possible. The selected sample sizes were chosen based upon practical limitations involving the number of events that might realistically be implemented. The stars do not fall strictly on the proportional allocation curve because of limitations associated with the sampling of integer-valued numbers of events.

These graphs indicate that, for temperature, it might be practical to achieve a standard deviation in the range of 0.6–1.0ºC, and that a much larger sample size would be needed to reduce the standard deviation below 0.5ºC. Similarly, standard deviations associated with estimates of mean relative humidity can possibly be achieved in the range of 1.5–2.5%, and realistic standard deviations associated with extinction coefficient might be in the range of 0.0045–0.0070 km^{-1}. Geographic variability with respect to many of these results is described later in this section.

Although the estimation of mean levels of parameters is likely to be a primary point of emphasis for many model-based results, the accurate estimation of extremes is also of significant importance. This issue was specifically addressed by investigating the precision associated with the estimation of the 90th percentiles of the evaluative parameters.

In contrast to the standard deviations associated with estimation of the mean, there is no closed-form solution to determine the variability associated with the estimation of 90th percentiles. Therefore, a Monte Carlo-type resampling approach was utilized to estimate these standard deviations. This specifically involved randomly selecting 200 artificial samples of actual data, each consisting of the required number of events, from the 20 seasonally defined strata. From each sample and at each site location, the 90th percentile of the parameter was estimated. The variance of the resulting collection of 200 estimates was averaged across sites, and the associated

standard deviation served as an estimate of the precision associated with the particular sample size. This exercise was repeated for sets of 30, 40, 50, and 60 events.

The graphs in Figure 17-8(a-c) illustrate the resulting standard deviations, along with the standard deviations associated with estimation of the mean using the indicated number of events. As would be expected, the standard deviations associated with estimation of the 90th percentile are higher than those associated with estimation of the mean of each parameter. This difference is most pronounced for extinction coefficient, with standard deviations for 90th percentile estimation being approximately three times those for mean estimation. For relative humidity, they are approximately twice as large. The difference is least pronounced for temperature, where the increase is less then 20%. In each case, there is only slight, gradual improvement in the precision of 90th percentile estimation for sample sizes of greater than 40 events.

The next step is to arrive at an appropriate number of events to be distributed among these 20 strata. Table 17-6 displays the standard deviation associated with the estimation of the annual mean of each evaluative parameter (both raw and seasonally adjusted) that would result from samples consisting of 30, 40, 50, and 60 events. For comparison purposes, these standard deviations represent average variances restricted to the RADM geographic domain. The table also displays the standard deviation associated with the aggregation of 30 3-day events in a sample stratified using the original RADM clusters. From this, it is clear that any number of events would be sufficient to provide improved resolution relative to the RADM scheme, in the context of the evaluative parameters reviewed here.

Although these results suggest that a 30-event sample would be sufficient to meet the objective of matching or exceeding the performance of the RADM approach with regard to estimation precision for these outcome parameters, other results displayed in this section demonstrate clear improvement in the precision associated with estimation of both means and extremes by moving from a 30-event sample to a 40-event sample. In addition, a 40-event set is needed to ensure equal precision to the RADM approach with regard to the estimation of wet deposition amounts. The RADM set of 30 events included 20 events from categories that were identified as “wet”, i.e., for which average wet SO_4^{2-} deposition exceeded the median for each cluster. In other words, 20 events were selected from the “wettest” 50% of all events.

This oversampling of wet events was originally pursued to ensure the adequate representation of those events that contributed most significantly to wet deposition, because the accurate characterization of wet deposition was the primary purpose of RADM at that time. However, concerns have since arisen that the disproportionate representation of these events may have introduced an overall bias with regard to the ambient concentrations of pollutants that are influenced by cloud cover and precipitation. In view of these concerns, and since wet deposition is not the primary focus for CMAQ, the oversampling of wet events was deemed inappropriate for present purposes. In the absence of this oversampling, it is still necessary to include a sufficient number of events to ensure that wet deposition is characterized as accurately here as in the RADM approach. This would require approximately 20 wet events, and the same median-

based definition of wet versus dry events implies that approximately 20 dry events should also be included. Therefore, a total of 40 events was deemed necessary to satisfy all of the objectives.

Table 17-6. Standard deviation[1] associated with estimation of the annual mean of the indicated parameter, using stratified sampling[2] with 20 strata and 30, 40, 50, or 60 events. *RADM comparative analysis.*

No. of Events	Temperature, deg. C		Relative Humidity, %		RH-Adjusted Extinction Coefficient, km^{-1} ($\times 10^3$)	
	20 Strata	Existing RADM Sample (30 Events)	20 Strata	Existing RADM Sample (30 Events)	20 Strata	Existing RADM Sample (30 Events)
	NOT SEASONALLY ADJUSTED					
30	0.96	1.78	2.34	2.58	6.65	8.28
40	0.82		2.01		5.60	
50	0.73		1.79		5.01	
60	0.66		1.63		4.59	
	SEASONALLY ADJUSTED					
30	0.74	0.82	2.32	2.41	6.64	8.23
40	0.63		1.99		5.60	
50	0.56		1.77		5.00	
60	0.51		1.61		4.58	

[1] The table entries reflect the standard deviation associated with the average station-specific variances (averaged across stations within the RADM geographic domain).

[2] "Existing RADM Sample" results reflect sampling of 3-day events from 1982-1985. Other results reflect sampling of 3-day events from 1984-1992.

Recalling that for 40 events the precision associated with the estimation of mean annual extinction coefficient (the primary evaluative parameter) was optimized using 20 seasonally defined strata (Table 17-5), a final plan was adopted for sampling 40 events from 20 strata (5 strata per season) using approximately proportional allocation.

Figure 17-9(a-d) displays star chart histograms of the 20 clusters defined as strata in this arrangement. Each chart illustrates the frequency of occurrence of 5-day events belonging to a given cluster, and the clusters themselves are ordered according to overall frequency of occurrence. The numbers arranged radially on each chart depict the number of events belonging to the cluster from each month of the year. Map-based graphs of mean wind vectors for day 3 of each cluster are contained in Figures 17-10 through 17-29.

In order to examine the impact of this scheme geographically (in the context of the precision associated with the estimation of mean levels of the evaluative parameters), we examined graphically (not shown), the standard deviation at each site location alongside the actual mean with which that standard deviation is associated. Similarly, analogous information associated with the estimation of 90th percentiles were examined. This examination confirmed the relative geographic uniformity with respect to the standard deviations, which serves as support for the use of "average" standard deviations in drawing conclusions throughout this section.

17.5.2 Selection of Stratified Sample of Events

A stratified sample of events was randomly selected from the 20 seasonally defined strata for the period 1984-1992. The sample was selected without replacement to ensure that no single day was selected into more than one five-day event, i.e., that there was no overlap between selected events. Systematic sampling (Cochran, 1977) was used within each stratum for which more than one event was to be selected. Specifically, all events assigned to the stratum were ordered chronologically, an event was selected near the beginning of that ordering, and subsequent events were selected to be evenly spaced throughout the remainder of the ordering. If k events were to be sampled from a cluster containing n events, to illustrate the simple case in which n/k is integer valued, the first event would be randomly selected from any of the chronologically first n/k events, and every (n/k)th subsequent event would be selected. The purpose of this approach was to ensure appropriate representation over the entire nine-year period.

Table 17-7 displays the total number of events belonging to each stratum, the number of events sampled, and the dates of the sampled events. These dates are the middle dates of the three-day events for which the model ultimately is to be run (i.e., the last three days of the sampled five-day event). This sample of 40 events includes representation from every month of the year, and from every year during the period 1984-1992. Table 17-8 illustrates this representation by displaying the number of events selected from each month and from each year.

Table 17-7. Stratum sizes, number of sampled events per stratum, and dates of events in sample. *Dates shown are for middle day of 3-day event.*

Stratum	Season	Total # of Days in Stratum, 1984-1992	Number of Sampled Events	Event Dates
1	Spring	292	3	12 March 1985, 08 May 1987, 27 March 1990
2	Summer	267	3	17 July 1985, 20 August 1987, 10 August 1990
3	Autumn	238	3	08 September 1986, 12 October 1988, 08 October 1991
4	Winter	210	3	04 January 1986, 15 December 1988, 02 December 1992
5	Spring	200	2	07 May 1984, 06 March 1990
6	Winter	188	2	03 January 1987, 07 January 1992
7	Spring	185	2	01 April 1986, 26 March 1991
8	Summer	171	2	05 August 1986, 29 June 1992
9	Summer	171	2	07 August 1984, 12 July 1989
10	Winter	170	2	18 January 1984, 25 January 1989
11	Autumn	168	2	18 October 1985, 12 September 1991
12	Autumn	150	2	17 November 1987, 14 September 1992
13	Winter	139	2	19 February 1985, 27 January 1990
14	Autumn	135	2	17 October 1988, 24 November 1991
15	Summer	129	2	03 July 1987, 09 July 1992
16	Autumn	128	2	25 November 1985, 07 November 1990
17	Winter	102	1	18 December 1989
18	Summer	90	1	22 July 1989
19	Spring	89	1	09 May 1990
20	Spring	62	1	30 April 1991

Table 17-8. Number of sampled events representing each month of the year and each year from the period 1984-1992.

Month	Number of Events in Sample	Year	Number of Events in Sample
January	6	1984	3
February	1	1985	5
March	4	1986	4
April	2	1987	5
May	3	1988	3
June	1	1989	4
July	5	1990	6
August	4	1991	5
September	3	1992	5
October	4		
November	4		
December	3		

17.6 Application and Evaluation

In this section, examples of the aggregation calculation for annual mean concentrations and total wet depositions, applicable to this sample of events, are provided. Following this example is a description of an evaluation exercise in which the aggregation calculation was carried out for light extinction coefficient, and the aggregated estimates were compared to the actual values based on data from all of the days in the period.

17.6.1 Application of the Aggregation Procedure

Aggregation calculations will be applied to model-based depositions and concentrations obtained for each sampled event, to achieve unbiased estimates for annual and seasonal means and other summary statistics within each grid cell. Since the goal of sampling from every defined stratum is achieved in this approach, these calculations are simplified in comparison to earlier aggregation methods (NAPAP, 1991). In essence, these aggregation calculations merely produce weighted means, totals, or other summary measures, from the sample of events.

To illustrate the aggregation approach, consider the estimation of a mean annual air concentration using model output for the 40 events selected above. These events represent 20 strata; denote these using the subscript i, i=1,2, . . . , 20. Let f_i denote the frequency of

occurrence associated with stratum *i*, i.e., the total number of 3-day events belonging to the stratum during the period 1984-1992. For an individual grid cell, also let

$$\bar{C}_{MODEL_i}$$

represent the mean model-based concentration associated with all sampled events from stratum *i*. Thus, for strata with a single sampled event, it is just the event mean concentration in the grid cell. For strata with two or three sampled events, it is the mean concentration for all of those events. Then the estimated annual mean air concentration is given by

$$Mean\ Air\ Concentration = \frac{\sum_{i=1}^{20} f_i \bar{C}_{MODEL_i}}{\sum_{i=1}^{20} f_i} \quad . \tag{17-2}$$

Estimates for most other parameters (e.g., dry depositions) and other summary statistics are calculated using similar methods. The calculation for wet deposition is different, primarily because the weighting is partially dictated by precipitation. Let

$$\bar{D}_{MODEL_i},\ \bar{P}_{MODEL_i},\ P_{MEAS_i}$$

represent the mean 3-day modeled deposition for sampled events in stratum *i*, the mean 3-day modeled precipitation for sampled events in stratum *i*, and the total measured precipitation accounted for by all events belonging to stratum *i*, respectively. Then the estimated total annual wet deposition is given by

$$Total\ Wet\ Deposition = \sum_{i=1}^{20} \left(\frac{\bar{D}_{MODEL_i}}{\bar{P}_{MODEL_i}} \right) \times P_{MEAS_i} \times \frac{1}{3 \times 9} \quad . \tag{17-3}$$

This expression can be thought of as a weighted sum in which the model-estimated wet concentration for a stratum is weighted by the total measured precipitation associated with the stratum. The final component of this expression is included to reflect the fact that each day is counted three times in the calculated sum (due to the use of 3-day events) and that the strata are defined over a nine-year period.

17.6.2 Evaluation

In order to determine the effectiveness of the aggregation technique and subsequent episode selection, comparisons were made between the observed mean b_{ext}s for the period 1984-1992 and the aggregated estimates of that mean using the stratified sample of events described in 17.5.2 and listed in Table 17-7. This preliminary evaluation, which includes simple regression analysis, is similar to that performed on RADM (Eder and LeDuc, 1996; Eder et al.,1996).

Results comparing the observed and aggregated mean b_{ext}s (based on Equation 17-2) are promising as seen in the scatterplot provide in Figure 17-30. The correlation between the 201 observed and aggregated mean b_{ext} was very high correlation ($r^2 = 0.988$). Estimates of the regression coefficients between the observed and aggregated mean b_{ext}s reveal an intercept value of -0.0012 that is not significantly different from zero ($\alpha = 0.05$). The slope (1.018), however; is significantly different from 1.0, ($\alpha = 0.05$) indicating a slight tendency for these particular episodes to provide an over-estimate of the expected mean b_{ext}. This slight, "apparent" bias is, however, well within the expected variability associated with the particular set of episodes in this stratified sample. To wit, selection of a different random set of episodes would just as likely result in a slight under-estimate of the expected mean b_{ext}.

Perhaps a better way to illustrate the effectiveness of this technique can be shown through an examination of the percent deviation in aggregate estimates of the mean b_{ext} (where the deviations are relative to the observed mean (aggregated b_{ext} - observed b_{ext} / b_{ext} observed). These percent deviations, which were calculated over the period 1984 -1992, are presented in Figure 17-31. For the most part, the deviations are within ± 10%, indicating excellent agreement between the actual mean b_{ext} and the aggregated estimates of the mean b_{ext}.

The slight over prediction tendency mentioned above appears to be somewhat spatially biased as also seen in Figure 17-31. As seen in the top of the figure, areas of generally positive deviations (aggregation approach yields a higher b_{ext}, hence lower visibilities than observed) appear to concentrate from east Texas into the southeastern states and again in the upper midwest between Minnesota and the Dakotas. The states of California and Idaho also exhibit positive deviations. Negative deviations, presented in the bottom of the Figure 17-31, tend to predominate from the northeast states into the Great Lake States and southwestward toward the states of Kansas, New Mexico and Arizona. This spatial dependence of the estimates is, once again, well within the expected variability. Selection of a different random set of episodes would likely result in a different pattern of positive and negative deviations, as there is a natural tendency for sites at close proximity to behave in a similar fashion.

The scatter plot in Figure 17-30 also reveals an increase in variance about the regression line starting at an observed mean b_{ext} of 0.085. Unlike the positive bias, discussed above, this increase in variance does not appear to be spatially biased, but rather exhibits a random distribution across the domain. This is represented in Figure 17-31 by the scattering of the larger biases (i.e. biases > 5.0%) evenly across the domain.

17.7 Summary and Discussion

The objective of this research was to develop a new aggregation approach and set of events to support model-based distributional estimates of air quality parameters (acidic deposition, air concentrations, and measures related to visibility) over the continental domain. The basic approach is to define meteorological categories that account for a significant proportion of the variability exhibited by these air quality parameters, as well as the particular transport mechanisms involved, so that source-attribution analyses are facilitated. This requires that categories be defined with an emphasis on wind flow parameters. To this end, the cluster analysis of zonal *u* and meridional *v* wind field components has been used to determine meteorologically representative categories.

The research described in this chapter was carried out in three phases:

- Phase 1: Various stratification schemes were evaluated and compared on different temporal and geographic scales to support the selection of a preferred general methodology. The selected methodology involved clustering of wind field data over the continental domain within each of the four seasons, and defining strata to be equivalent to the resulting clusters. This methodology demonstrated superior relative efficiency compared to methods defined on an annual time frame or on a regional scale for estimates involving the evaluative meteorological parameters, and is designed to support seasonal estimation with both simplicity and precision.

- Phase 2: Determinations were made regarding appropriate numbers of clusters and events to support sampling using the general methodology selected in Phase 1. The resulting scheme involved a total of 20 clusters (5 per season), and 40 events, defined over the time period 1984-1992. This scheme affords superior precision to previous approaches for estimates involving the evaluative meteorological parameters, supports approximately equivalent representation of wet events to those approaches without oversampling, and provides adequate resolution of wind field patterns that characterize transport.

- Phase 3: A stratified sample of events was selected under approximate proportional allocation, using systematic sampling within strata, in accordance with the scheme determined in Phase 2. This sample was successfully evaluated through a comparison of aggregated estimates of the mean b_{ext} to the actual mean b_{ext}, revealing a high level of agreement, although there was a slight tendency of the aggregation and episode selection technique to over-estimate the expected mean b_{ext}.

The goal of this research was to categorize many years worth of meteorological patterns into a few classes. This represents a very ambitious goal, and it should not be surprising that there is substantial variability associated with the wind vectors assigned to individual clusters. Nevertheless, the results described above suggest that the approach achieves a reasonable characterization of frontal passage scenarios and leads to clusters that explain variation in the evaluative meteorological parameters used in this analysis (temperature, relative humidity, and

visibility), and therefore can be used to achieve improved estimates of these parameters relative to estimates obtained from simple random sampling. Moreover, the constrained definition of distinct seasonally-based clusters brings further improvement to the ability of the clusters to explain the variation in these parameters, and therefore leads to more precise estimates associated with them. The evaluative parameters were selected for their known associations with many air quality parameters of interest, thus suggesting that the clusters should also be effective in defining strata from which events can be selected to estimate those air quality parameters.

17.8 References

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

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

Cochran, W.G., 1977. *Sampling Techniques*. Wiley & Sons, New York.

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

Eder, B. K., J. M. Davis and P. Bloomfield, 1994. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. *Journal of Applied Meteorology*, **33**, 1182-1199.

Eder, B. K., S.K. LeDuc and F. Vestal, 1996. Aggregation of selected RADM simulations to estimate annual ambient air concentrations of fine particulate matter. Ninth Joint Conference on Applications of Air pollution Meteorology with the A&WMA, Jan.28-Feb. 2, Atlanta, GA.

Eder, B. K. and S.K. LeDuc, 1996. Can selected RADM simulations be aggregated to estimate annual concentrations of fine particulate matter? Proceedings of the International Specialty Conference on the Measurement of Toxic and Related Air Pollutants, May 7 - 9, Research Triangle Park, NC.

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

Fernau, M.E. and P.J. Samson, 1990b. Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part II: Precipitation patterns and pollutant deposition. *Journal of Applied Meteorology*, **29**, 751-761.

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. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne and D. Joseph, 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc*., **77**, 437-471.

NAPAP, 1991. *National Acid Precipitation Assessment Program 1990 Integrated Assessment Report*, National Acid Precipitation Assessment Program, 722 Jackson Place NW, Washington, D.C.

SAS Institute Inc., 1989. *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1*. SAS Institute Inc., Cary, NC.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236-244.

This chapter is taken from *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System,* edited by D. W. Byun and J. K. S. Ching, 1999.

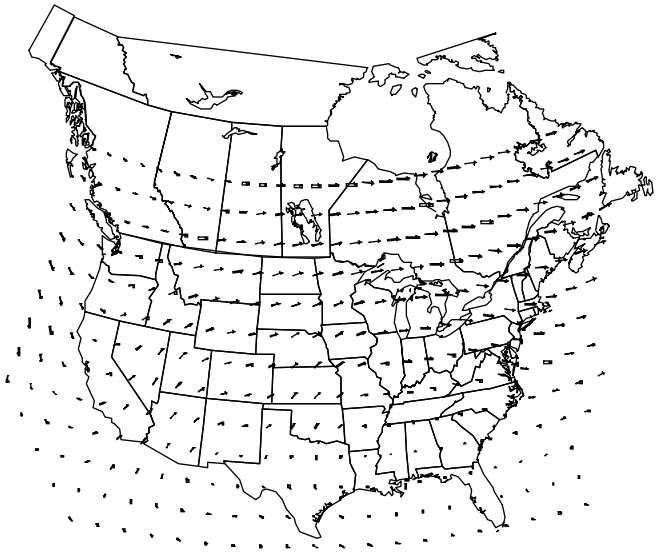

Fig.17.1 (a) Mean wind vectors for day 1 of annually defined cluster 1 (of 30).

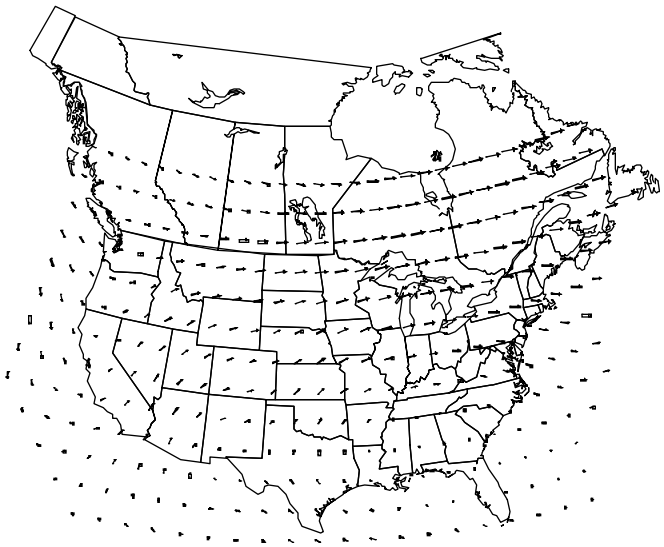

Fig.17.1 (b) Mean wind vectors for day 3 of annually defined cluster 1 (of 30).

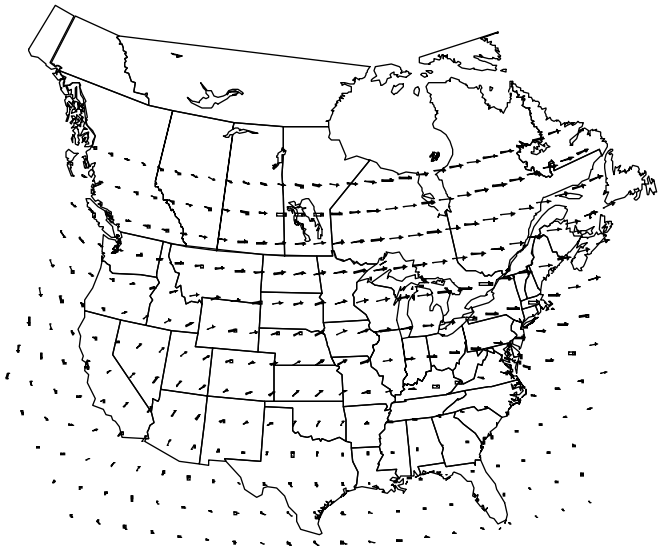

Fig.17.1 (c) Mean wind vectors for day 5 of annually defined cluster 1 (of 30).