Development of an Aggregation and Episode Selection Scheme to Support the Models-3 Community Multiscale Air Quality Model

RICHARD D. COHN

Analytical Sciences, Inc., Durham, North Carolina

BRIAN K. EDER,* SHARON K. LEDUC,* AND ROBIN L. DENNIS*

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, Research Triangle Park, North Carolina

(Manuscript received 8 November 1999, in final form 10 March 2000)

ABSTRACT

The development of an episode selection and aggregation approach, designed to support distributional estimation for use with the Models-3 Community Multiscale Air Quality (CMAQ) model, is described. The approach utilized cluster analysis of the 700-hPa east–west and north–south wind field components over the time period of 1984–92 to define homogeneous meteorological clusters. Alternative schemes were compared using relative efficiencies and meteorological considerations. An optimal scheme was defined to include 20 clusters (five per season), and a stratified sample of 40 events was selected from the 20 clusters using a systematic sampling technique. The light-extinction coefficient, which provides a measure of visibility, was selected as the primary evaluative parameter for two reasons. First, this parameter can serve as a surrogate for particulate matter with diameter of less than 2.5 μm, for which few observational data exist. Second, of the air quality parameters simulated by CMAQ, this visibility parameter has one of the most spatially and temporally comprehensive observational datasets. Results suggest that the approach reasonably characterizes synoptic-scale flow patterns and leads to strata that explain the variation in extinction coefficient and other parameters (temperature and relative humidity) used in this analysis, and therefore the approach can be used to achieve improved estimates of these parameters relative to estimates obtained using other methods. Moreover, defining seasonally based clusters further improves the ability of the clusters to explain the variation in these parameters and therefore leads to more precise estimates.

1. Introduction

In support of studies mandated by the 1990 Clean Air Act Amendments, the Models-3 Community Multiscale Air Quality (CMAQ) model (Byun and Ching 1999) is used to estimate various pollutant air concentrations and depositions as related to specified levels of emissions. Congressionally mandated air pollution assessment studies conducted by the U.S. Environmental Protection Agency (EPA) involve calculation of benefits, for various effects, associated with seasonally and annually focused emissions reductions. These studies require CMAQ-based distributional estimates for a variety of pollutants based upon representative meteorological conditions. The relevant pollutants include ozone concentrations, for crop effects (U.S. EPA 1999); acidic and nutrient deposition, for ecosystem effects (U.S. EPA 1995, 1999); and fine particle concentrations for health effects (U.S. EPA 1998, 1999) and visibility degradation (U.S. EPA 1999; Malm et al. 1994; Chestnut and Dennis 1997).

CMAQ would ideally simulate atmospheric chemical processes associated with meteorological conditions occurring on a daily basis over several years. However, for logistical and financial reasons, it is not currently feasible to execute CMAQ over an extended time period such as a full year. Run time, even on fast computers, can approach 50% of the simulation time (i.e., a 24-h simulation could require close to 12 h of central processing unit time). 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

* On assignment to the National Exposure Research Laboratory, U.S. Environmental Protection Agency.

Corresponding author address: Brian K. Eder, ASMD/ARL, Mail Drop 80, Research Triangle Park, NC 27711.
E-mail: eder@hpcc.epa.gov

of the research described in this paper is to develop such an aggregation approach and to evaluate its effectiveness.

Relevant background information regarding this activity, a precise statement of the current objectives, and an overview of the strategy are provided in this section. Section 2 describes the development of the general approach for selecting events. The refinement of a specific approach is detailed in section 3. Sample selection and evaluation are presented in section 4, and section 5 provides a summary and discussion.

a. Background

The basic problem of developing representative meteorological categories has been explored by other researchers for a variety of purposes (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; Chang et al. 1987) to estimate annual averages.

The approach of Brook et al. involved four major components. Cluster analysis of wind fields was used to determine meteorologically representative categories. Determining 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.

b. Objective

The objective described in this paper is the development of an episode selection and aggregation approach that supports model-based annual and seasonal air quality estimates, in consideration of the broad applicability envisioned for CMAQ both with respect to air quality parameters and geographic representation. RADM was primarily designed to address issues involving acidic deposition, but CMAQ addresses a more diverse collection of air quality parameters (i.e., tropospheric ozone, acid deposition, visibility, and particulate matter) under a different set of priorities. Seasonal as well as annual timescales are of interest. In addition, CMAQ will employ a continental domain that is significantly larger than the geographic area employed by RADM (eastern United States and Canada). Extending the domain to a continental domain is extremely ambitious for episode selection and aggregation. The development of an approach that accommodates this larger continental domain is particularly challenging.

c. Strategy overview

The methods employed herein 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 such optimality can be measured.

A specific goal is to define meteorological categories that account for a large 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 with modifications to reflect the differences inherent in the current objectives. The common element is the cluster analysis of zonal u and meridional v wind components to define meteorological categories.

It might be argued that the categories should be defined directly using the air quality parameters that are of interest (e.g., ozone or fine particles). 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 objective requires that meteorological categories be defined with an emphasis on wind flow parameters. Indeed, characterizations of basic wind field patterns describe synoptic-scale flow, along with all of the meteorological and air quality properties typically associated with them.

The basic strategy used for the selection of events to support aggregation-based estimation is described in the steps outlined below.

1) Different approaches were explored for the cluster analysis of wind components; each of these approaches resulted in the definition of a set of clusters (strata) of meteorologically similar events. 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 paper, stratum and cluster refer to the same collections of events, because the clusters defined in the analysis are ultimately used as the strata for sampling purposes.
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 complex sampling schemes in comparison with that of a simple random sampling scheme.
3) An appropriate number of clusters to retain was determined in combination with an acceptable number of events that would be necessary to achieve sufficiently small variances in the estimates. The deter-

mination was based on estimated standard deviations associated with several alternative formulations and on other considerations explained below.
4) A stratified sample of events was randomly selected from the clusters defined in step 3.

2. General approach to episode selection

The development of a general approach to episode selection is described in this section. This approach involves the cluster analysis of u and v wind components to define meteorological categories.

a. Description of wind data

To accommodate the continental domain and to achieve adequate spatial resolution, the cluster analysis involved data at 336 grid nodes with 2.5° horizontal spacing, as obtained from the National Centers for Environmental Prediction–National Center for Atmospheric Research 40-Year Reanalysis Project (Kalnay et al. 1996). In this analysis, 700-hPa 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 meteorological conditions, to support an emphasis on continental regions for which air quality modeling is focused. Graphical illustrations of this domain are referenced later in this section.

b. Cluster analysis of wind data

Cluster analysis, in the current 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-yr period (1984–95, defined to maximize the amount of available data) was considered in the exploratory cluster analyses, later refined to a 9-yr base period (1984–92, defined to minimize air quality trends stemming from the implementation of controls during the mid-1990s associated with the 1991 Clean Air Act Amendments) for the final clustering upon which episode selection was based. Because CMAQ is actually run for a 5-day period for each event (the first two days establish initial conditions, and model results from days 3–5 are saved as a ''3-day event''), 5-day periods were clustered rather than 3-day periods as in Brook et al. (1995a). 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 2016 elements (the 2 u and v components $\times$ 336 grid nodes $\times$ 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 to 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 3 days) of the squared deviations of each u and v component from the cluster mean u and v. In the final cluster analysis, using 1984–92 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. This approach was necessary for computational efficiency; ancillary analyses demonstrated that the results are insensitive to the choice of consecutive versus overlapping periods or the use of different starting points.

It is useful to consider preliminary results for a set of 30 clusters initially defined using year-round data from 1984 to 1995 (i.e., cluster analysis of daily wind field data across this set of years, without regard to season). Figures 1a–c illustrate mean wind vectors for a representative cluster that accounted for 3.65% of all 5-day events during this time period (ranking ninth in overall prevalence). Most of these events occurred between the months of October and April and were characterized by northwesterly winds associated with a large-scale trough moving through the central and eastern portion of the domain.

Although these maps are effective in illustrating average behavior associated with each cluster, they do not indicate the variability inherent in the clusters. Figure 2 shows the mean wind vectors for day 3 of the same cluster on a grid that only includes alternating grid nodes. The map contains groups of small dots; each 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 the cluster.

The dots surrounding each mean wind vector arrowhead appear in a somewhat circular pattern, and the 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, which emphasizes the ambitious nature of this endeavor. In essence, the goal is

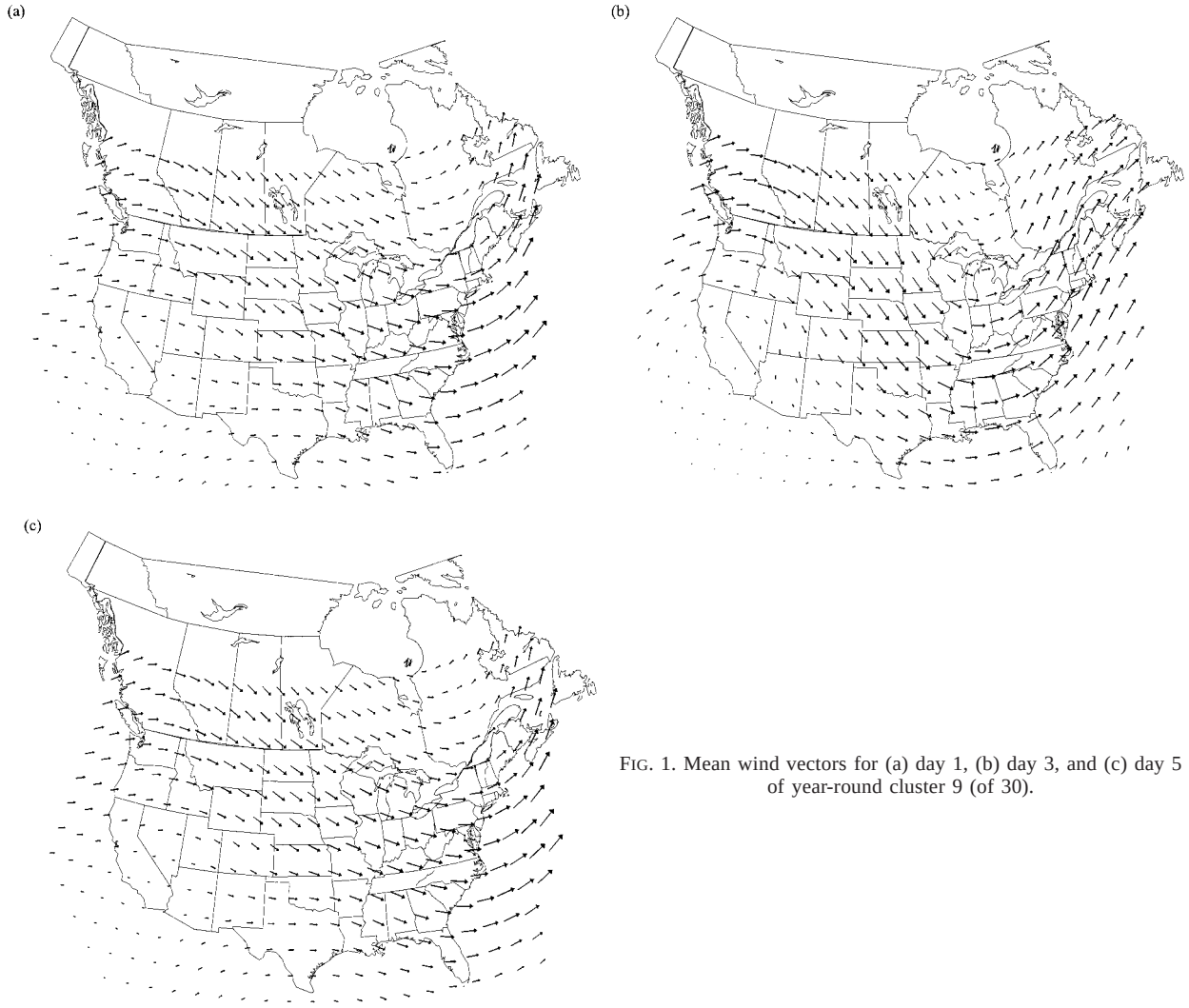

FIG. 1. Mean wind vectors for (a) day 1, (b) day 3, and (c) day 5 of year-round cluster 9 (of 30).

to categorize many years 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 represent simultaneously 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; however, Ward's method of cluster analysis ensures that the within-cluster variability is minimized.

Preliminary results also demonstrated that, although defined using year-round data, the cluster frequencies reveal definite seasonal tendencies. That is, clusters do not occur randomly throughout the year, but tend to occur more frequently within specific seasons. Thus, the year-round clustering procedure successfully identifies and discriminates wind field patterns that are associated with seasonally distinct meteorological classes. This property was also observed by Brook et al. (1995a). Furthermore, the two most prevalent clusters heavily emphasized 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, given that the wind fields are expected to be less variable in the summer. However, seasonal differences in meteorological and atmospheric chemical conditions are important in explaining the variability exhibited by the air quality parameters of interest. Adding more warm season clusters would provide improved resolution in this regard.

c. Clustering and evaluation methods

The preliminary clustering described above clarifies the motivation for investigating seasonally distinct clustering. An added dimension to the problem is that the number of clusters to be retained clearly affects the amount of variability in the wind fields that can be accounted for by the clusters. To gain an understanding

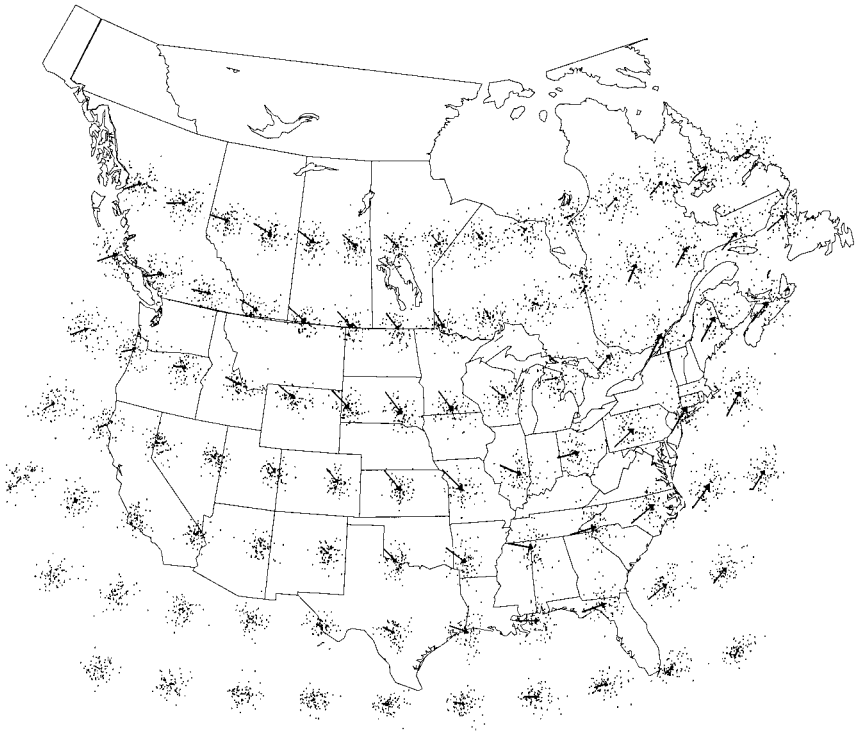

FIG. 2. Mean and distribution of wind vectors for day 3 of cluster 9 (of 30).

of the importance of such considerations as they relate to estimation of the meteorological parameters used as evaluative tools in this analysis, seasonal alternatives were explored in several combinations.

Several seasonal variations of the u and v wind vector clustering were investigated. These were selected to investigate patterns and properties of cluster definitions. They include 1) year-round strata with variations in the number of strata; 2) seasonally defined strata using warm (April–September) and cold (October–March) seasons; 3) seasonally defined strata using summer (June, July, and August), winter (December, January, and February), and transitional (spring and autumn combined) seasons; and 4) seasonally defined strata using summer, winter, spring (March, April, and May), and autumn (September, October, and November) seasons.

Although wind flow parameters were used to define the meteorological categories, other parameters were used to refine aspects of the episode selection methodology and to evaluate the effectiveness of the approach. These parameters include surface visibility (as represented by the light-extinction coefficient b_{ext}), temperature, and relative humidity. These meteorological parameters serve to characterize frontal passages and related transport and chemical characteristics, and appropriately resolved data are available for them. For consistency with the typical application of CMAQ, these parameters were determined from days 3–5 of each 5-day period for purposes of evaluating the methodology.

In this analysis, emphasis is placed on b_{ext}, which provides a surrogate measure for fine particulate matter. Elevated concentrations of fine particles are of great concern to EPA because they have been linked to detrimental health effects and visibility degradation. An evaluation of the applicability of the approach to other pollutants simulated by CMAQ, including ozone and acid deposition, is currently underway and will be reported elsewhere. It must be recognized that this approach constitutes an indirect evaluative tool, in the sense that the effectiveness of the approach is not directly measured as it relates to atmospheric transport or to specific air quality parameters, both of which are primary outcomes.

For the alternative stratification schemes, preliminary testing was performed by examining the uncertainty associated with the use of cluster-based stratified sampling to estimate these parameters, specifically the annual mean of daily noontime levels of visibility, temperature, and relative humidity. Visibility was specifically expressed as b_{ext} (km^{-1}), omitting observations with precipitation and observations with relative humidity greater than 90%. A correction factor, which normalized all of the b_{ext} values to a common relative humidity value of 60%, was also applied to the data to compensate for the influence of humidity.

The light-extinction coefficient is often used to characterize visibility, although in general it has limited ability to predict human visual perception. The visual range

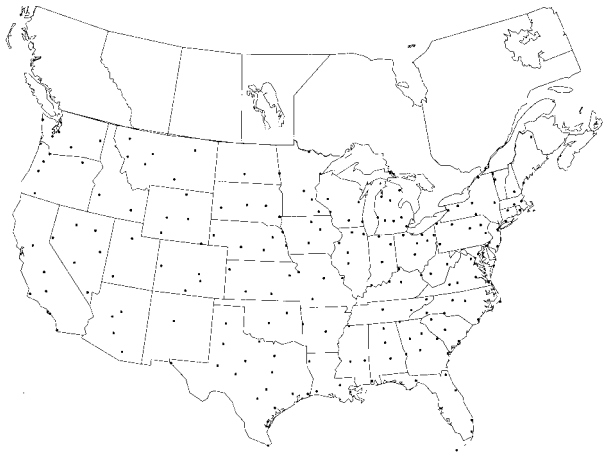

FIG. 3. Site locations for 201 meteorological parameters used in the evaluation.

v_r (km) can be estimated from b_{ext} by using the Koschmieder equation:

$$v_r = \frac{3.91}{b_{ext}}. \quad (1)$$

These meteorological parameters were taken from 201 locations in the continental United States (National Weather Service first-order stations) for which b_{ext} coverage was at least 99%, as illustrated in Fig. 3. Allowances were made for missing observations that were associated with precipitation so as not to bias the inclusion of sites toward drier climates.

d. Sampling considerations

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 method is potentially much more efficient than equal allocation, in the sense that it leads to much more precise (i.e., lower variance) estimates.

The average relative efficiency associated with the estimation of mean annual light extinction, temperature, and relative humidity, using each alternative stratification scheme, was used to compare the schemes. 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.

Specifically, at each location and for each scheme, the variance of an aggregation-based estimate (Cochran 1977) of the annual mean was determined using 1) a stratified sample with proportional allocation across strata, and 2) simple random sampling, with a total sample size consisting of the same numbers of events in each case. Then at each location, the ratio of the simple random sampling variance to the variance associated with the stratified sampling design was calculated and expressed as the relative efficiency of the design. Last, those relative efficiencies were averaged across sites to indicate the overall efficiency of each scheme.

This approach can be illustrated with a hypothetical example. 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 entire period 1984–92, disregarding the clusters. Suppose that the standard deviation associated with that estimate is 1.5°C, so that a 95% confidence interval based on a normal distribution would yield the estimated mean $\pm$ 2.94°C ($=1.96 \times 1.5$). Next, suppose that the mean is instead estimated as a weighted average from 30 3-day events that are selected from strata using proportional allocation. Suppose that the standard deviation associated with that estimate is 1.0°C, as compared with 1.5°C from simple random sampling. These standard deviations translate to variances of 1.0 and 2.25, respectively. Thus, for this hypothetical location, the relative efficiency of stratified sampling with proportional allocation is $2.25/1.0 = 2.25$. In general, stratified sampling using proportional allocation is more efficient than simple random sampling in the sense that it leads to lower variances and therefore tighter confidence intervals.

e. Cluster analysis results

Table 1 presents mean relative efficiencies associated with annual means of the daily noontime extinction coefficient, temperature, and relative humidity, as estimated using aggregation approaches based upon the various schemes described above. Relative efficiencies reported in this section are valid for any number of events that might be selected from the indicated number of strata, because relative efficiency is invariant to sample size under proportional allocation. The first six rows in Table 1 illustrate relative efficiencies associated with various numbers of year-round strata (i.e., clusters emerging from cluster analyses of daily wind field data from 1984 to 1995 without regard to season). The relative efficiency for mean temperature is consistently

TABLE 1. Mean relative efficiency associated with estimation of the annual (1984–95) mean of the indicated parameter, using various stratified sampling approaches using proportional allocation relative to simple random sampling.

Row	No. of strata and method of definition	Extinction coefficient	Temperature	Relative humidity
1.	5, year-round	1.12	2.41	1.18
2.	10, year-round	1.14	2.50	1.22
3.	20, year-round	1.16	2.83	1.26
4.	30, year-round	1.17	2.86	1.28
5.	60, year-round	1.19	3.05	1.32
6.	90, year-round	1.20	3.10	1.33
7.	10, seasonal (5 warm, 5 cold)	1.13	2.94	1.26
8.	20, seasonal (10 warm, 10 cold)	1.16	3.54	1.35
9.	30, seasonal (15 warm, 15 cold)	1.17	3.60	1.37
10.	60, seasonal (30 warm, 30 cold)	1.20	3.89	1.39
11.	20, seasonal (6 summer, 6 winter, 7 transitional)	1.18	3.91	1.31
12.	30, seasonal (10 summer, 10 winter, 10 transitional)	1.20	4.12	1.35
13.	20, seasonal (5 summer, 5 winter, 10 transitional)	1.19	4.06	1.33
14.	30, seasonal (8 summer, 8 winter, 14 transitional)	1.20	4.17	1.36
15.	20, seasonal (5 summer, 5 winter, 5 spring, 5 autumn)	1.18	3.86	1.36
16.	30, seasonal (8 summer, 8 winter, 7 spring, 7 autumn)	1.19	3.88	1.38

highest, followed by relative humidity, and then extinction coefficient. This result implies that the meteorological clusters are most effective in distinguishing between events with regard to temperature and are least effective in distinguishing extinction coefficient. Furthermore, as expected, the efficiency increases as the number of strata increases.

Rows 7–10 of Table 1 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. For temperature and relative humidity, stratification using these seasonally defined clusters consistently yields improved efficiency over stratification using the same number of year-round clusters. Although the year-round 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 comparison, an analysis basing stratum definitions upon season alone (with no further wind field–based clustering) yielded mean relative efficiencies of 1.12, 2.93, and 1.21 for extinction coefficient, temperature, and relative humidity, respectively (Cohn et al. 1999).

Results associated with other seasonal stratification schemes are illustrated in the next six rows of the table. Rows 11 and 12 correspond to approximately equal numbers of clusters divided among summer, winter, and transitional (spring and autumn combined) seasons, rows 13 and 14 correspond to equal numbers of clusters divided between summer and winter with approximately twice as many transitional season clusters, and rows 15 and 16 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 (20 or 30) investigated in other seasonal and year-round analyses. Stratification schemes based on three or four seasons offer improved efficiency in the estimation of mean temperature in comparison with two-season and year-round stratification schemes with comparable total numbers of strata. They also demonstrate slight but uniform improvement for extinction coefficient but mixed results for relative humidity.

Additional analyses for all three evaluation variables (Cohn et al. 1999) demonstrated that, when using non-proportional allocation, the relative gains in efficiency associated with seasonally defined clusters versus year-round clusters are larger than the potential gains realized by seasonal versus year-round clustering when using proportional allocation. Thus, in departing from proportional allocation (which is not precisely achievable in practice because of the impossibility of sampling fractional numbers of events), seasonally defined clusters are likely to afford improved efficiency over year-round clusters. For example, under proportional allocation, the relative efficiency for extinction coefficient improved by 2% in moving from 20 year-round clusters to 20 clusters divided equally among four seasons, whereas it improved by 33% for the same comparison under equal allocation. Related analyses further demonstrated that, through seasonal clustering and approximate proportional allocation, improved efficiency can be realized relative to the stratification scheme used in the aggregation of RADM output, with respect to the outcome measures used here. When the extension to the continental domain is considered, the efficacy of seasonal clustering with approximate proportional allocation is further elevated.

3. Refinement of the sampling approach

Because three- and four-season stratification schemes were found to be better than other schemes, a four-season scheme was selected after considering differ-

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

After selection of this general approach, refinements were needed to determine an appropriate number of strata and to determine an adequate number of events for sampling. This analysis was limited to 1984–92 (during which time emission changes were minimal) to facilitate later testing with CMAQ. Baseline meteorological conditions for use in air quality modeling would be chosen from this period.

a. Determination of the number of strata

The precision associated with the estimation of annual means of the evaluative parameters was investigated for various combinations of up to 28 seasonally defined strata. More strata would make it impractical to ensure that every stratum is sampled, that is, that there are no strata that go unrepresented in the final set of events. At the estimation stage, an unsampled stratum would simply be combined with another stratum to which it is most similar. In practice, this combination was achieved by constraining the selection to collections of strata that were adequate to support the sampling of at least one event from each stratum.

For a realistic range of 30–60 events, precision did not improve with increasing numbers of strata beyond a threshold of 20 strata. For example, in a 16-stratum, 40-event sample, the standard deviation of the estimated annual mean extinction coefficient was 5.71×10^3 km^{-1}; this value was reduced to 5.63×10^3 km^{-1} in a 20-stratum sample, but then it increased to 5.70×10^3 km^{-1} when 24 strata were used. Reduced precision for increasing numbers of strata results from the scheme's inability to satisfy proportional allocation.

The coefficient of determination R^2, that is, the proportion of variance in the u and v wind field components accounted for by the stratification, was also considered. In an arrangement consisting of 20 seasonally defined strata (five per season), the strata account for 18.3% of the total variance exhibited by the wind field data. By definition of the clustering method, a higher R^2 would result if the same number of year-round strata were used, because there would be more similarity in wind fields among some strata defined for different seasons than among strata defined using a cluster analysis without regard to season. This is the price that is accepted in exchange for enhanced clarity in the resolution of other characteristics that arises from seasonal clustering. Indeed, 20 strata defined year-round account for 23.9% of the total variance. Although a comparable R^2 might be realized for the 19 year-round clusters that were defined in the original RADM scheme, these clusters were not all represented in that sample, so that the effective R^2 would have been reduced somewhat from 23.9%, and we suspect that it is reasonably close to 18.3%. In consideration of these results, 20 strata (five per season) were determined to constitute an appropriately sized set.

Note that this analysis does not adhere to traditional rules of thumb regarding the determination of appropriate numbers of clusters to retain in a cluster analysis. 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 very smoothly. There is no ''correct'' number of clusters after which the relative importance of additional clusters drops markedly. 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.

b. Determination of number of events

The next step is to arrive at an appropriate number of events to be distributed among these 20 strata. 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 approach 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 Monte Carlo approach was repeated for sets of 30, 40, 50, and 60 events.

Figure 4 illustrates, as expected, that 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.

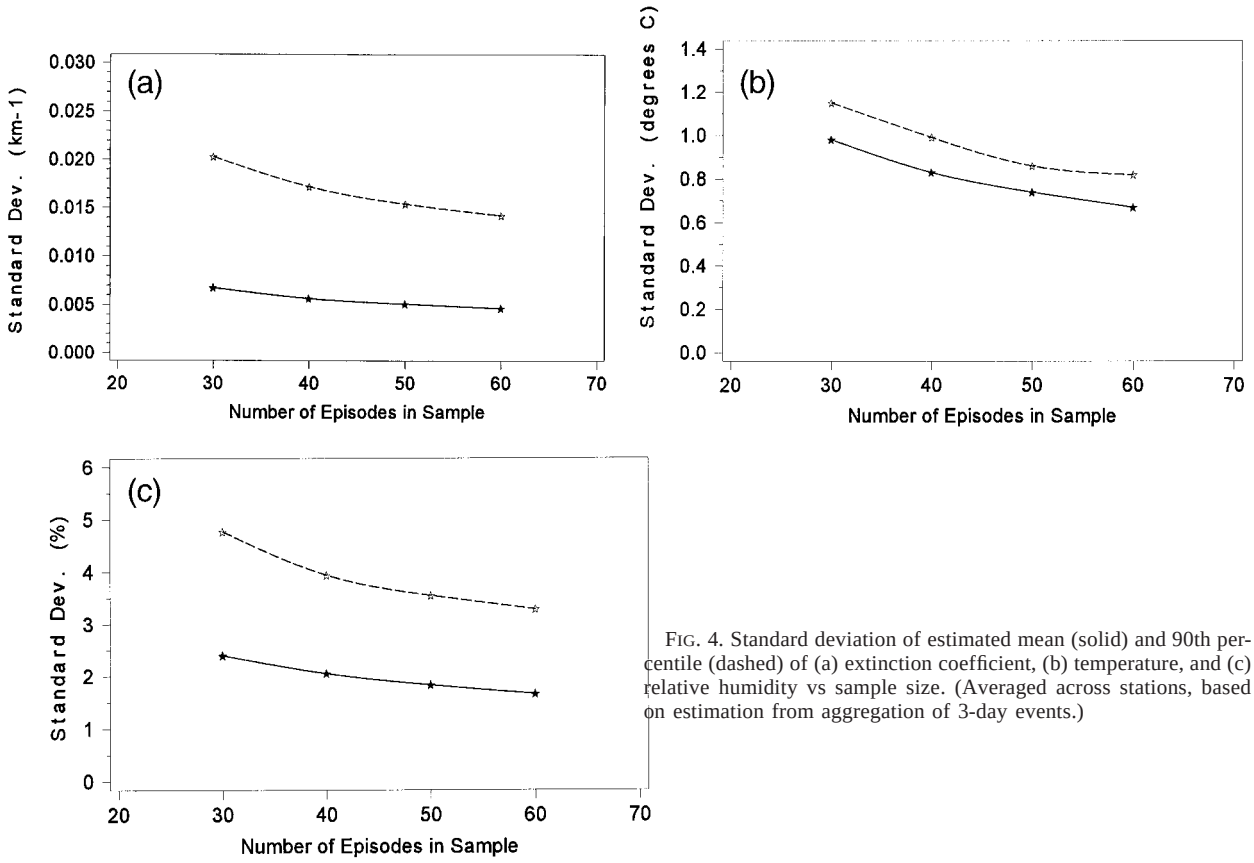

FIG. 4. Standard deviation of estimated mean (solid) and 90th percentile (dashed) of (a) extinction coefficient, (b) temperature, and (c) relative humidity vs sample size. (Averaged across stations, based on estimation from aggregation of 3-day events.)

The difference is least pronounced for temperature, for which the increase is less than 20%. For each parameter, precision of the 90th-percentile estimation improves only slightly for sample sizes greater than 40 events.

Table 2 displays the standard deviation associated with the estimation of the annual mean of each evaluative parameter that would result from samples consisting of 30, 40, 50, and 60 events. For comparison, these standard deviations represent average variances restricted to the RADM geographic domain and are listed beside the standard deviation associated with the aggregation of 30 3-day events in a sample stratified using the original RADM clusters. These results collectively 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. Beyond 40 events, improvements diminished and computational constraints increased; therefore, the selection of 40 events was deemed to be an acceptable compromise.

c. Characterization of the sampling scheme

In consideration of these results, a sampling scheme consisting of 40 events, selected from 20 seasonally defined strata (five strata from each of four seasons),

TABLE 2. Comparison with RADM sample of the standard deviation associated with estimation of the annual mean of the indicated parameter, using stratified sampling with 20 strata and 30, 40, 50, or 60 events. Table entries reflect the standard deviation associated with the average station-specific variances (averaged across stations within the RADM geographic domain). ''RADM sample'' results reflect sampling of 30 3-day events from 1982 to 1985. Other results reflect sampling of 3-day events from 1984 to 1992.

	Extinction coefficient, km^{-1} ($\times 10^3$)		Temperature, °C		Relative humidity, %	
No. of events	20 seasonally defined strata	RADM sample	20 seasonally defined strata	RADM sample	20 seasonally defined strata	RADM sample
30	6.65		0.96		2.34	
40	5.60	8.28	0.82	1.78	2.01	2.58
50	5.01		0.73		1.79	
60	4.59		0.66		1.63	

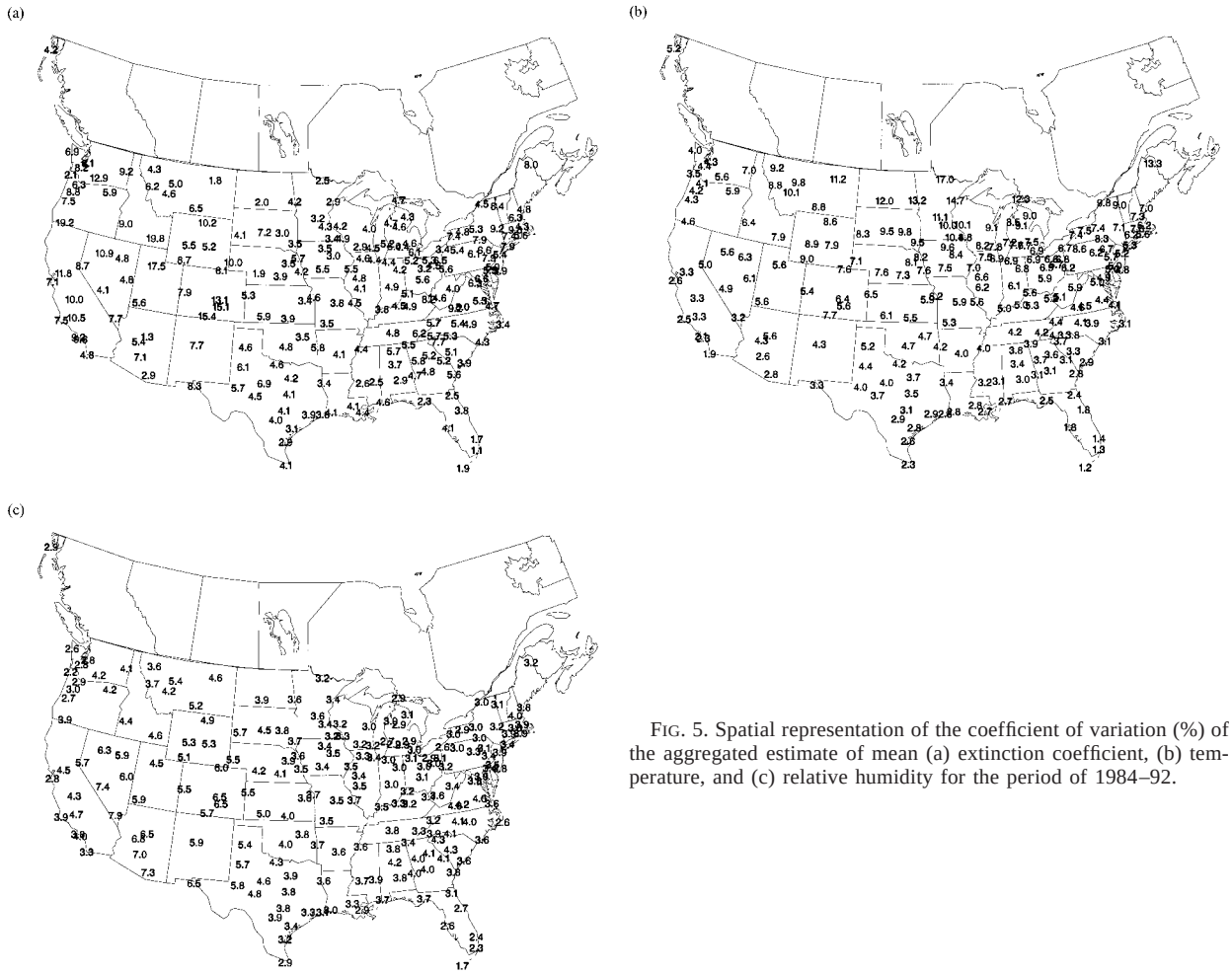

FIG. 5. Spatial representation of the coefficient of variation (%) of the aggregated estimate of mean (a) extinction coefficient, (b) temperature, and (c) relative humidity for the period of 1984–92.

was adopted for use. Figure 5 geographically illustrates the precision associated with the estimation of mean levels of the evaluative parameters under this scheme, expressed as coefficient of variation (CV, the standard deviation as a percent of the mean), at each site location. For all three evaluative parameters, the CVs are usually in the range of 3%–7% and are nearly always less than 10%.

Table 3 displays frequencies of occurrence of the 20 clusters defined as strata in this arrangement. Each row 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 (representing spring months) accounted for 8.88% of all 5-day events between 1984 and 1992, those from cluster 2 (representing summer months) accounted for 8.12% of all events, and so on. The remaining entries on each row 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 the 5-day events from each cluster are contained in the appendix.

Figure 6 graphically demonstrates the manner in which the clusters account for much of the variability in the extinction coefficient, by illustrating the different distributions of daily noontime extinction coefficient among the clusters at four representative sites across the United States: Wilkes-Barre, Pennsylvania; Charlotte, North Carolina; Glasgow, Montana; and Sacramento, California. The geographic differences among the four sites demonstrate the challenging nature of the work, because the goal is a single stratification scheme that is simultaneously applicable in each of these locations. The seasons themselves clearly account for much of the variation in extinction coefficient, but the clusters also account for major differences within seasons. For example, the median extinction coefficient for Charlotte in cluster 14 (Fig. A14) is 0.11 km^{-1} (Fig. 6b), which corresponds to a visual range of 35.5 km. This cluster, which occurs during autumn, is associated with a deep trough centered over the Great Lakes. The resulting strong westerly flow in the Charlotte area is conducive to ventilation and therefore good visibility. Conversely, the median extinction coefficient for cluster 12 is 0.15

TABLE 3. Monthly frequencies of occurrence of the 20 seasonally defined clusters, 1984–92. Number of events belonging to each cluster from each month, and relative frequency (percentage of all events) of each cluster.

Cluster	Relative frequency (%)	Winter			Spring			Summer			Autumn		
		Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
1	8.88				127	75	90						
2	8.12							55	89	123			
3	7.24										107	96	35
4	6.42	55	86	70									
5	6.08				58	77	65						
6	5.78	64	56	70									
7	5.63				38	53	94						
8	5.20							44	53	74			
9	5.20							65	61	45			
10	5.20	68	59	44									
11	5.11										51	68	49
12	4.56										91	46	13
13	4.23	48	53	38									
14	4.11										21	56	58
15	3.92							63	42	24			
16	3.89										0	13	115
17	3.10	44	25	33									
18	2.74							43	34	13			
19	2.71				41	39	9						
20	1.89				15	26	21						

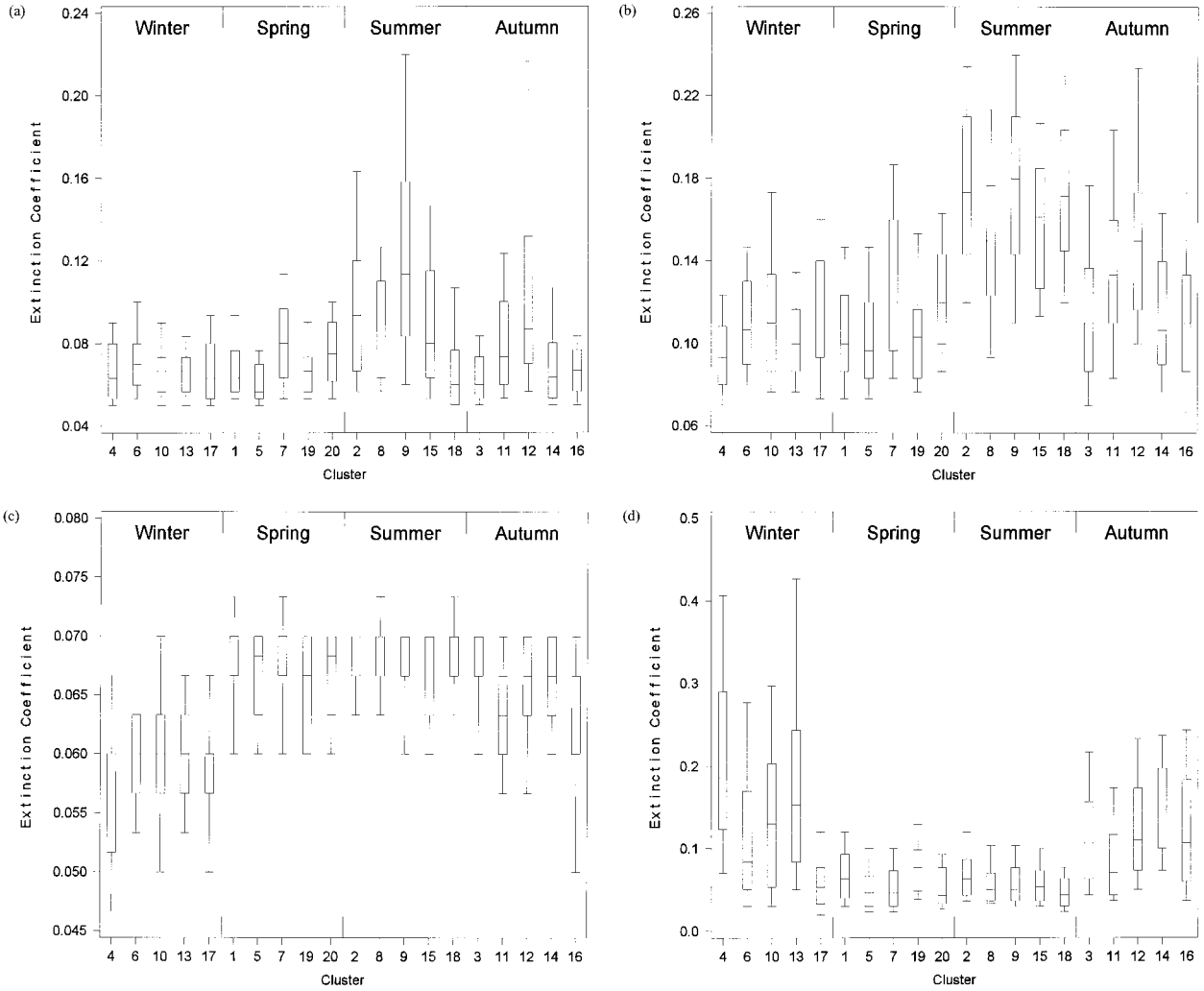

FIG. 6. Distribution of 3-day mean extinction coefficient, km^{-1} ($\times 10^3$), by cluster at (a) Wilkes-Barre, PA, (b) Charlotte, NC, (c) Glasgow, MT, and (d) Sacramento, CA. Boxplots indicate 10th, 25th, 50th, 75th, and 90th percentiles.

km^{-1} (26.1 km). This cluster, which also occurs during autumn, is associated with a weak anticyclone centered over the southeastern states (Fig. A12). The 700-hPa flow is very weak, resulting in stagnating conditions that are conducive to poorer visibility.

4. Sample selection and evaluation

A stratified sample of events was randomly selected from the 20 seasonally defined strata for the period of 1984–92. The sample was selected without replacement to ensure that no single day was selected into more than one 5-day event, that is, 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 9-yr period.

Table 4 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 3-day events for which the model ultimately is to be executed (i.e., the last 3 days of the sampled 5-day event). This sample of 40 events includes representation from every month of the year (range 1–6 events per month) and from every year during the period of 1984–92 (range 3–6 events per year).

In practice, aggregation calculations are applied to model-based estimates obtained for each sampled event

TABLE 4. 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 events in stratum	No. of sampled events	Sampled event dates
1	Spring	292	3	12 Mar 1985, 8 May 1987, 27 Mar 1990
2	Summer	267	3	17 Jul 1985, 20 Aug 1987, 10 Aug 1990
3	Autumn	238	3	8 Sep 1986, 12 Oct 1988, 8 Oct 1991
4	Winter	211	3	4 Jan 1986, 15 Dec 1988, 2 Dec 1992
5	Spring	200	2	7 May 1984, 6 Mar 1990
6	Winter	190	2	3 Jan 1987, 7 Jan 1992
7	Spring	185	2	1 Apr 1986, 26 Mar 1991
8	Summer	171	2	5 Aug 1986, 29 Jun 1992
9	Summer	171	2	7 Aug 1984, 12 Jul 1989
10	Winter	171	2	18 Jan 1984, 25 Jan 1989
11	Autumn	168	2	18 Oct 1985, 12 Sep 1991
12	Autumn	150	2	17 Nov 1987, 14 Sep 1992
13	Winter	139	2	19 Feb 1985, 27 Jan 1990
14	Autumn	135	2	17 Oct 1988, 24 Nov 1991
15	Summer	129	2	3 Jul 1987, 9 Jul 1992
16	Autumn	128	2	25 Nov 1985, 7 Nov 1990
17	Winter	102	1	18 Dec 1989
18	Summer	90	1	22 Jul 1989
19	Spring	89	1	9 May 1990
20	Spring	62	1	30 Apr 1991

to achieve unbiased estimates for annual and seasonal means and other summary statistics within each grid cell. Because the goal of sampling from every defined stratum is achieved in this approach, these calculations are simplified in comparison with 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 concentration using model output for the 40 events selected above. These events represent 20 strata. Let f_i denote the frequency of occurrence associated with stratum i, that is, the total number of 3-day events belonging to the stratum during the period of 1984–92. For an individual grid cell, also let $\overline{C}_{\mathrm{MODEL}_i}$ 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 3-day event mean concentration in the grid cell. For strata with two or three sampled events, it is the mean concentration for all of those 3-day events. Then, an unbiased estimate of the annual mean concentration is given by

$$\overline{C} = \frac{\sum_{i=1}^{20} f_i \overline{C}_{\mathrm{MODEL}_i}}{\sum_{i=1}^{20} f_i}. \tag{2}$$

The sampling variance of this estimate is given by

$$\mathrm{Var}(\overline{C}) = \sum_{i=1}^{20} \left(\frac{f_i}{\sum_{i=1}^{20} f_i} \right)^2 \left(\frac{S_i^2}{n_i} \right), \tag{3}$$

where S_i denotes the variance of all concentrations within stratum i, and n_i denotes the number of events actually sampled from stratum i. (Figure 5 was based on calculations of this form.) Under strictly proportional allocation, n_i is proportional to f_i, and, in a sample of 40 events,

$$\mathrm{Var}(\overline{C}) = \frac{1}{40} \sum_{i=1}^{20} \left(\frac{f_i}{\sum_{i=1}^{20} f_i} \right) S_i^2. \tag{4}$$

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

$$\overline{D}_{\mathrm{MODEL}_i}, \qquad \overline{P}_{\mathrm{MODEL}_i}, \qquad P_{\mathrm{OBS}_i}$$

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 observed precipitation accounted for by all events belonging to stratum i, respectively. Then, an estimate of the total annual wet deposition is given by

$$D = \sum_{i=1}^{20} \left(\frac{\overline{D}_{\mathrm{MODEL}_i}}{\overline{P}_{\mathrm{MODEL}_i}} \right) P_{\mathrm{OBS}_i} \frac{1}{3 \times 9}. \tag{5}$$

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 observed 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 (because of the use of 3-day events) and that the strata are defined over a 9-yr period.

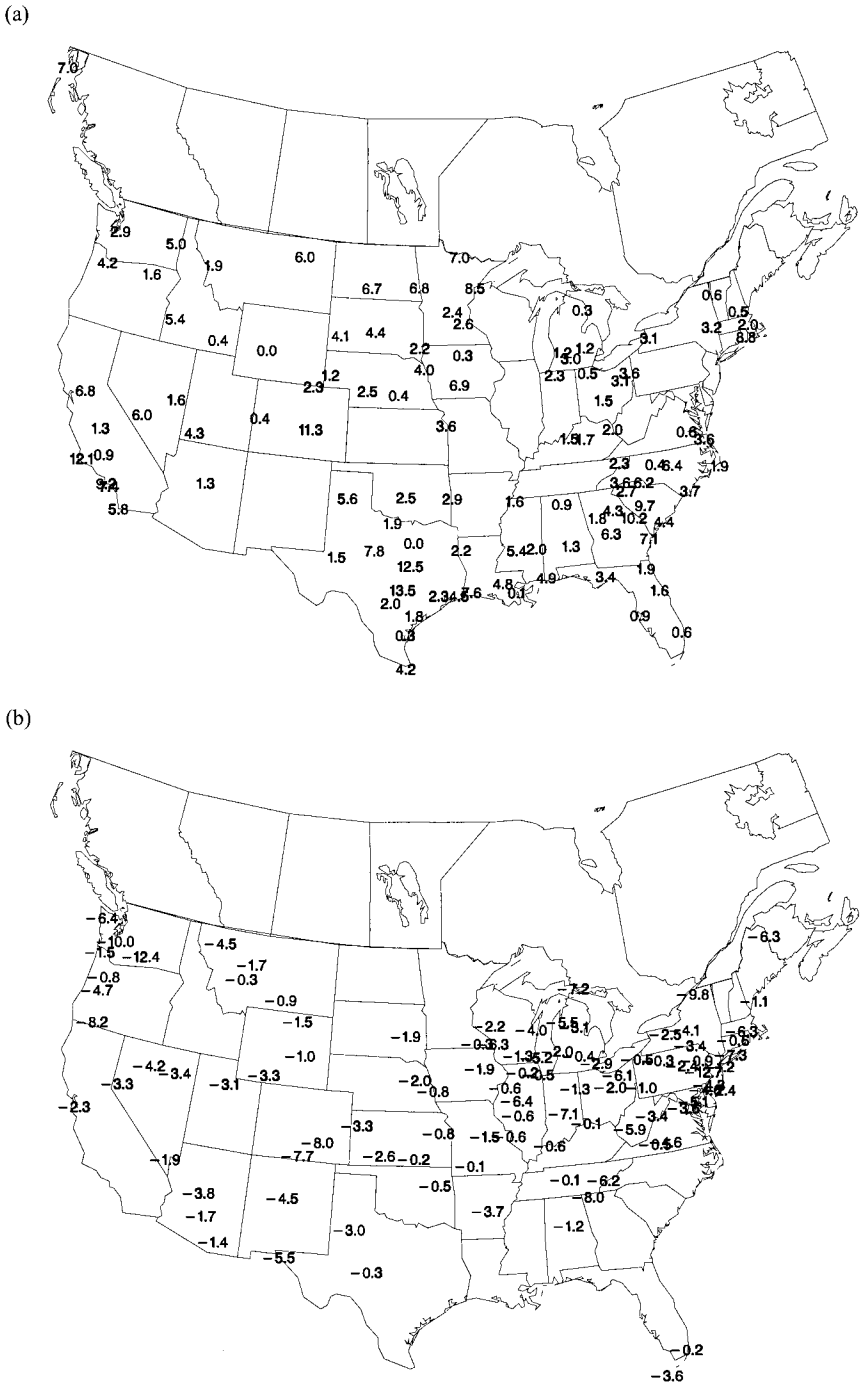

FIG. 7. Spatial representation of the (a) positive and (b) negative deviations of the aggregated estimates of the mean extinction coefficient for the period of 1984–92. [Deviations (%) are relative to the observed mean: (aggregate − observed)/observed.]

To illustrate the aggregation technique, comparisons were made between the observed mean b_{ext} for the period of 1984–92 and the aggregated estimates of that mean using the stratified sample of events listed in Table 4. This evaluation is similar to previous ones performed on RADM (Eder and LeDuc 1996; Eder et al. 1996). Specifically, the percent deviations in aggregated estimates of the mean extinction coefficient (where the deviations are relative to the observed mean) were calculated over the period of 1984–92 and are presented in Fig. 7. For the most part, the deviations are within ±10% and are completely consistent with the coeffi-

cients of variation presented in Fig. 5, indicating excellent agreement between the actual mean and the aggregated estimates of the mean. The strong spatial dependencies in the b_{ext} data are reflected in the tendency for areas of under- and over-simulation to be clustered together geographically; however, the figure also illustrates the scattering of the larger deviations (i.e., deviations $>5.0\%$ in absolute value) evenly across the domain.

5. Summary

The objective of this research was to develop a new aggregation approach and set of events to support CMAQ-based distributional estimates of fine particles and other air quality parameters over the continental domain. The primary strategy involved categorizing many years of meteorological patterns into a few classes. This categorization represents a very ambitious goal, and, not surprisingly, the wind vectors assigned to individual clusters varied substantially. Nevertheless, the results suggest that the approach reasonably characterizes flow patterns associated with synoptic-scale patterns and leads to strata that explain variation in extinction coefficient, temperature, and relative humidity.

Defining seasonally based clusters further improved the ability of the clusters to explain the variation in these parameters and therefore led to more precise estimates. The final scheme included 20 clusters (five per season), and stratified systematic sampling was used to select a sample of 40 events from the 20 clusters. The approach performed better than simple random sampling: relative efficiencies were 1.18 for extinction coefficient, 3.86 for temperature, and 1.36 for relative humidity. A basic aggregation technique was also illustrated for the selected sample of events and revealed aggregated estimates of b_{ext} that fall generally within $\pm 10\%$ of the observed mean b_{ext} for the period of 1984–92, consistent with independently estimated variability.

Future work will investigate the ability of the aggregation and episode selection scheme to replicate actual b_{ext} on finer temporal and spatial scales, in order to accommodate various applications of CMAQ. For instance, will aggregated estimates of b_{ext} for an individual year such as 1988 (a meteorologically anomalous year) still fall within $\pm 10\%$ of the observed mean, or will the estimates deteriorate? Likewise, will the approach, which was developed on a continental scale, exacerbate the spatial dependencies noted in section 4 when it is applied to various regional scales? These concerns will be addressed in the future as specific CMAQ simulations are planned and performed.

Acknowledgments. The authors thank Dr. Jeffrey Brook of Environment Canada for his expertise and guidance during the progression of this research; Ms. Renee Jaramillo of Analytical Sciences, Inc., for development of graphical tools; and the two anonymous referees for their critical reviews and insightful suggestions. This work was supported by the U.S. Environmental Protection Agency under General Services Administration Contract GS-35F-4750G. This document has been reviewed and approved by the U.S. Environmental Protection Agency for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

APPENDIX

Mean Wind Vectors from Each Cluster

Mean wind vectors for the middle (third) day of the 5-day events from each cluster are depicted in Figs. A1–A20.

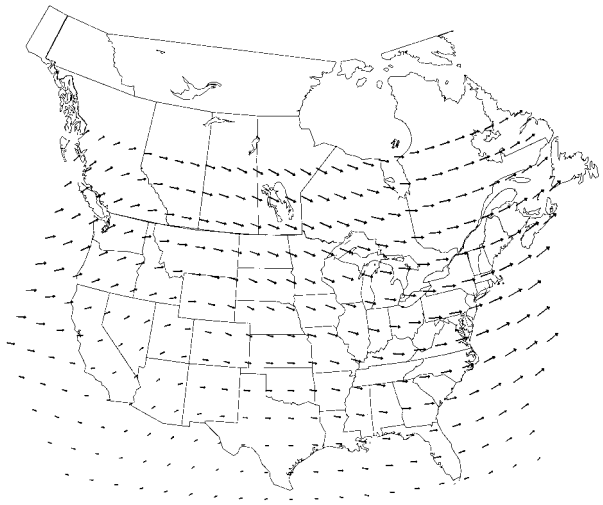

FIG. A1. Cluster 1, spring.

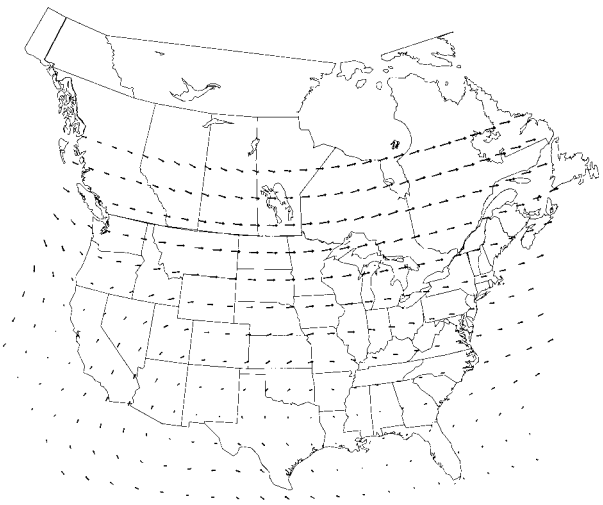

FIG. A2. Cluster 2, summer.

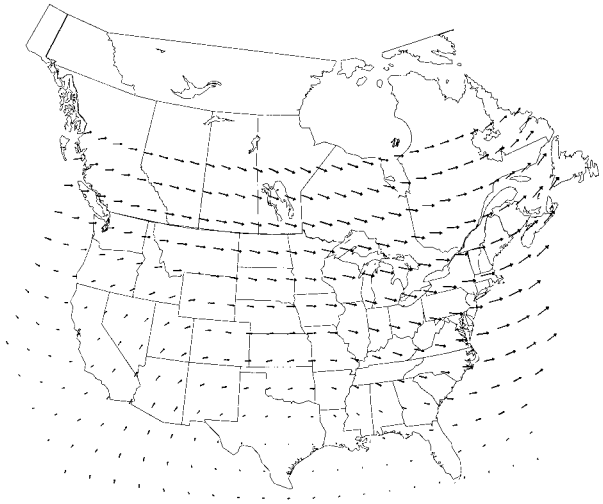

FIG. A3. Cluster 3, autumn.

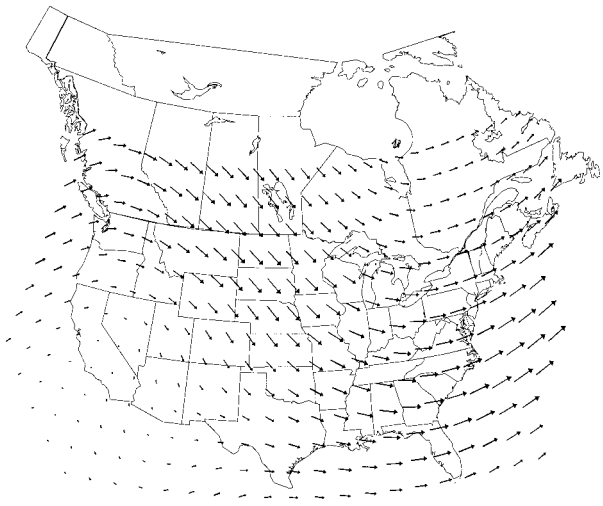

FIG. A4. Cluster 4, winter.

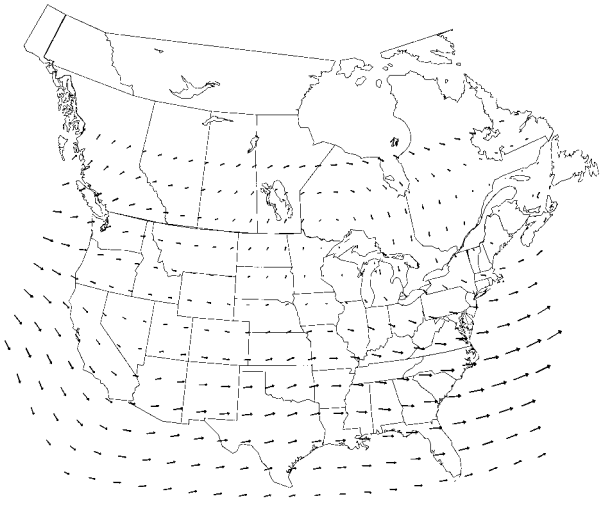

FIG. A5. Cluster 5, spring.

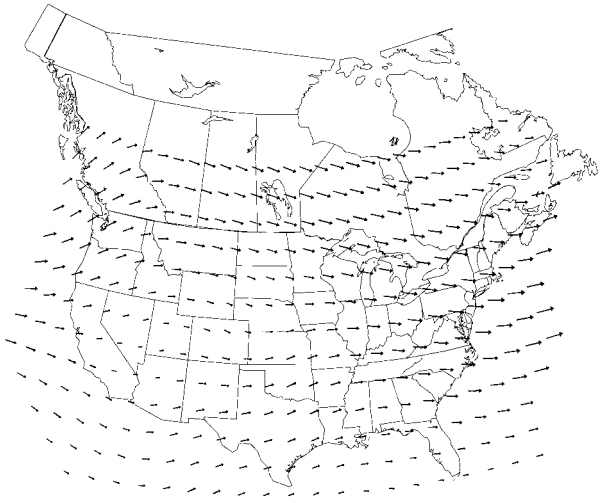

FIG. A6. Cluster 6, winter.

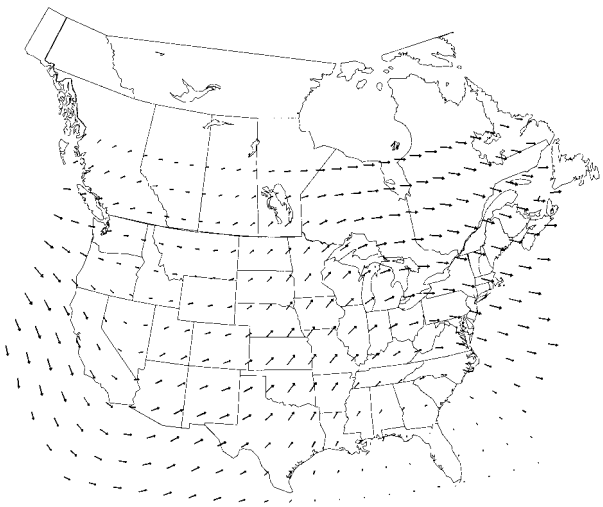

FIG. A7. Cluster 7, spring.

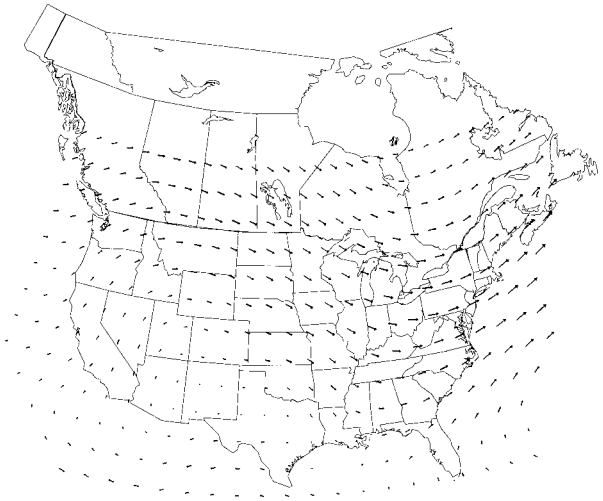

FIG. A8. Cluster 8, summer.

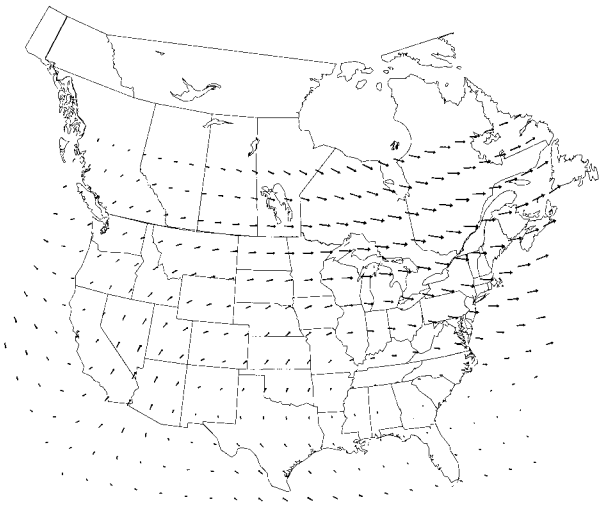

FIG. A9. Cluster 9, summer.

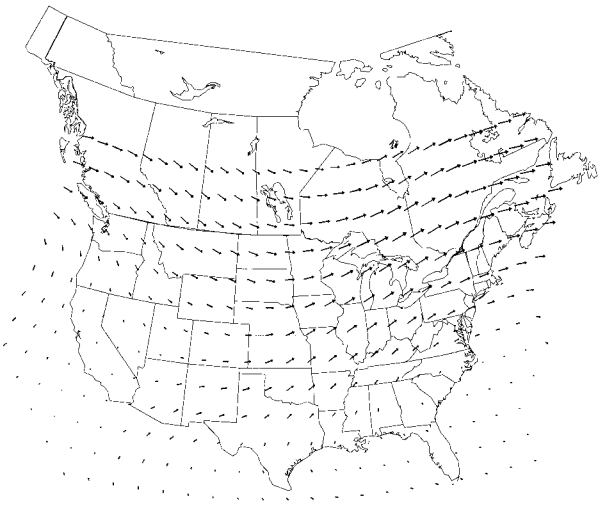

FIG. A12. Cluster 12, autumn.

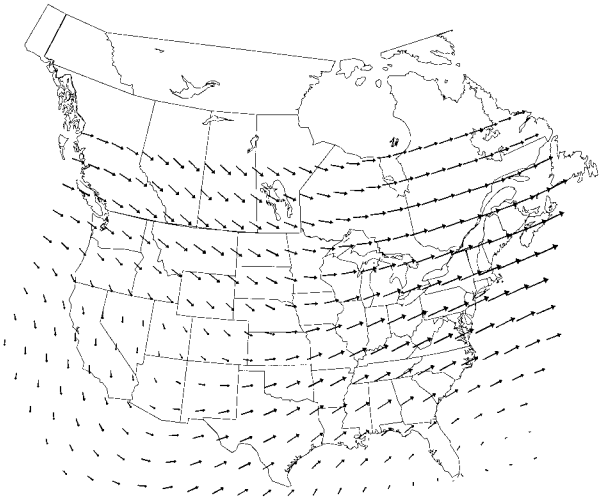

FIG. A10. Cluster 10, winter.

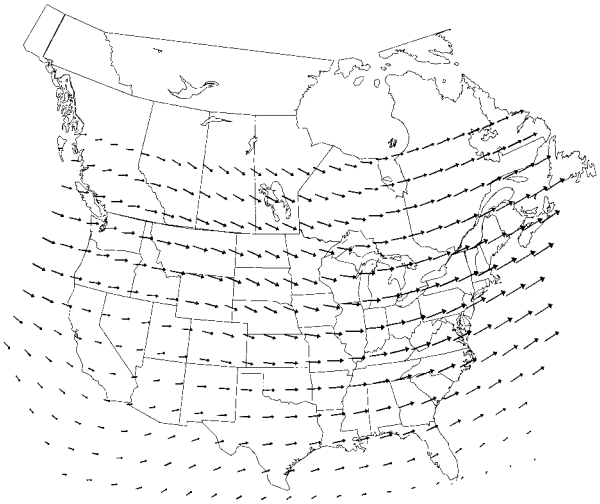

FIG. A13. Cluster 13, winter.

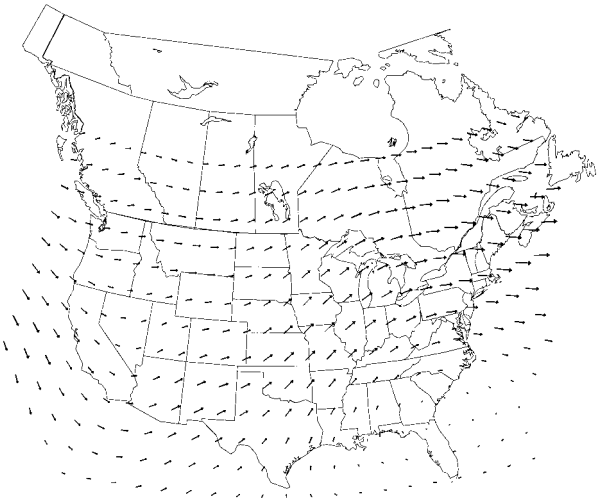

FIG. A11. Cluster 11, autumn.

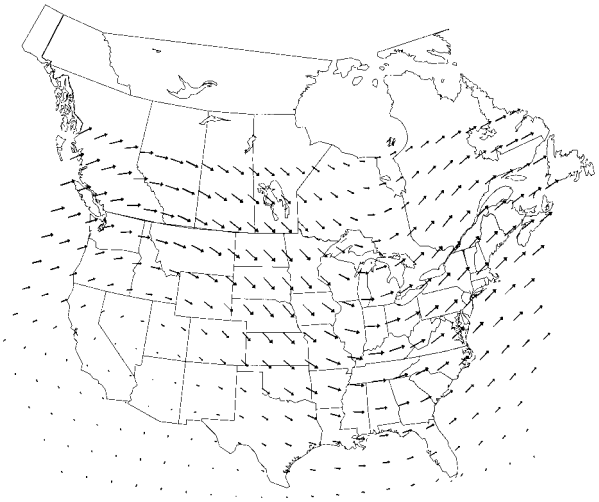

FIG. A14. Cluster 14, autumn.

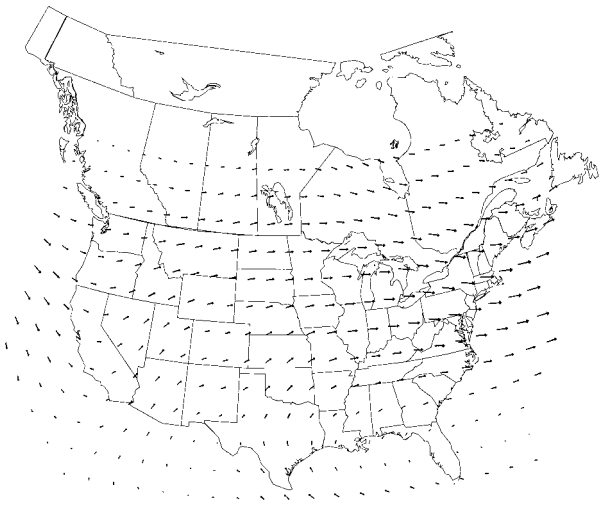

FIG. A15. Cluster 15, summer.

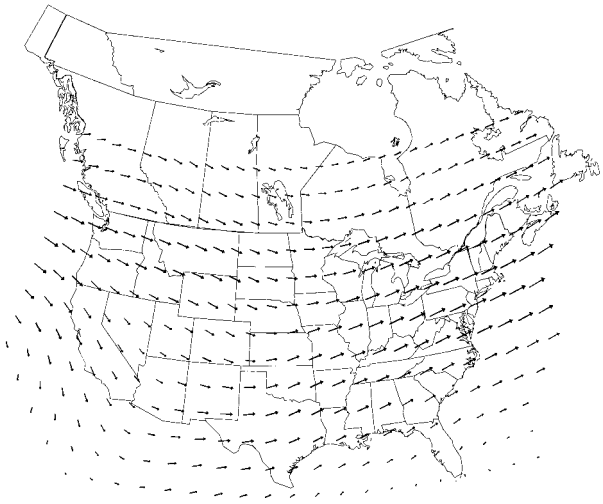

FIG. A16. Cluster 16, autumn.

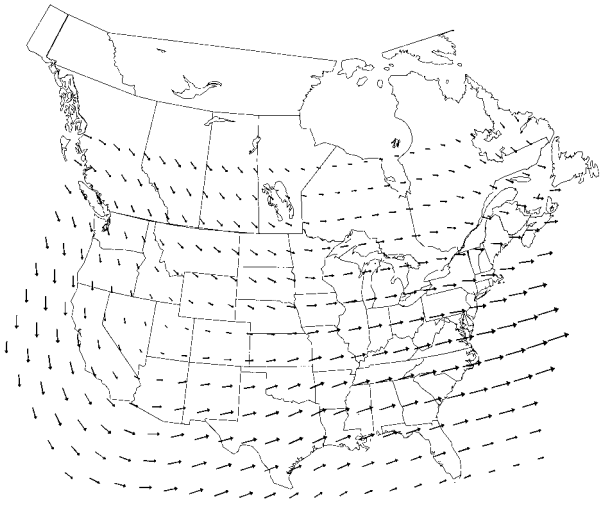

FIG. A17. Cluster 17, winter.

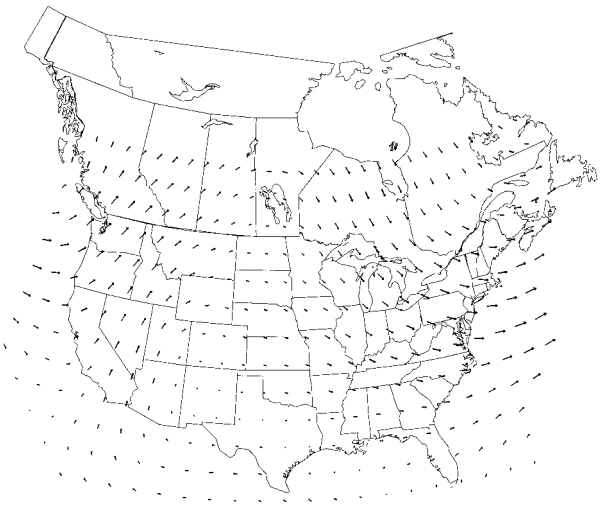

FIG. A18. Cluster 18, summer.

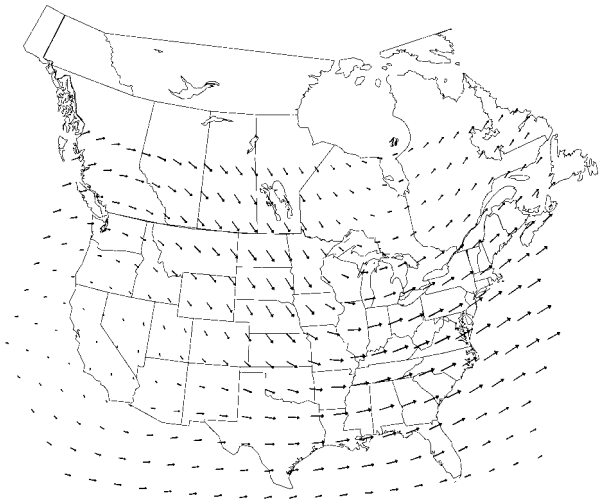

FIG. A19. Cluster 19, spring.

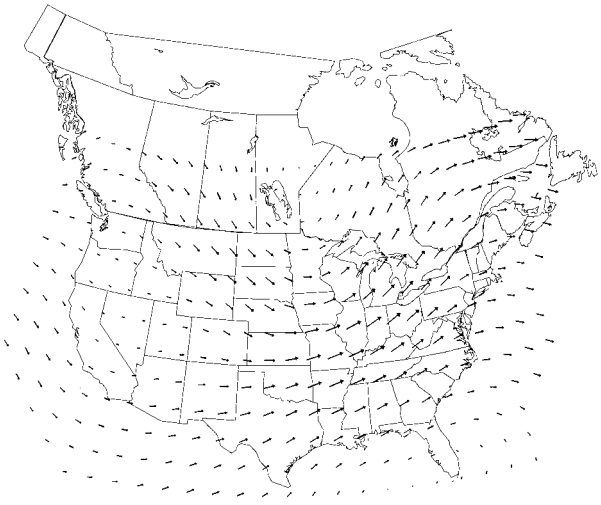

FIG. A20. Cluster 20, spring.

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. *J. Appl. Meteor.*, **34,** 297–325.

——, ——, and ——, 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. *J. Appl. Meteor.*, **34,** 326–339.

Byun, D. W., and J. K. S. Ching, Eds., 1999: Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system. EPA Tech. Rep. EPA-600/R-99/030, 757 pp. [Available from EPA/ORD, Washington, DC 20460.]

Chang, J. S., R. A. Brost, I. S. A. Isaksen, S. Madronich, P. Middleton, W. R. Stockwell, and C. J. Walcek, 1987: A three-dimensional Eulerian acid deposition model: Physical concepts and formulations. *J. Geophys. Res.*, **92,** 14 681–14 700.

Chestnut, L. G., and R. L. Dennis, 1997: Economic benefits of improvements in visibility: Acid rain provisions of the 1990 Clean Air Act Amendments. *J. Air Waste Manage. Assoc.*, **47,** 395–402.

Cochran, W. G., 1977: *Sampling Techniques.* John Wiley and Sons, 428 pp.

Cohn, R. D., B. K. Eder, and S. K. LeDuc, 1999: An aggregation and episode selection scheme designed to support Models-3 CMAQ. Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system, D. W. Byun, and J. K. S. Ching, Eds., EPA Tech. Rep. EPA-600/R-99/030, 17.1–17.65. [Available from EPA/ORD, Washington, DC 20460.]

Davis, R. E., and L. S. Kalkstein, 1990: Development of an automated spatial synoptic climatological classification. *Int. J. Climatol.*, **10,** 769–794.

Eder, B. K., and S. K. LeDuc, 1996: Can selected RADM simulations be aggregated to estimate annual concentrations of fine particulate matter? Preprints, *Int. Specialty Conf. on the Measurement of Toxic and Related Air Pollutants*, Research Triangle Park, NC, Air and Waste Management Association, 732–739.

——, J. M. Davis, and P. Bloomfield, 1994: An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. *J. Appl. Meteor.*, **33,** 1182–1199.

——, S. K. LeDuc, and F. Vestal, 1996: Aggregation of selected RADM simulations to estimate annual ambient air concentrations of fine particulate matter. Preprints, *Ninth Joint Conf. on Applications of Air Pollution Meteorology*, Atlanta, GA, Amer. Meteor. Soc., 390–392.

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. *J. Appl. Meteor.*, **29,** 735–750.

——, and ——, 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. *J. Appl. Meteor.*, **29,** 751–761.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77,** 437–471.

Malm, W. C., J. Trijonis, J. Sisler, M. Pitchford, and R. L. Dennis, 1994: Assessing the effect of SO_2 emission changes on visibility. *Atmos. Environ.*, **28,** 1023–1034.

NAPAP, 1991: National Acid Precipitation Assessment Program 1990 Integrated Assessment Report, 520 pp. [Available from the National Acid Precipitation Assessment Program, 722 Jackson Place NW, Washington, DC 20503.]

U.S. EPA, 1995: Acid deposition standard feasibility study report to Congress. U.S. EPA Rep. EPA 430-R-95-001a. [Available from EPA/OAR, Washington, DC 20460.]

——, 1998: Regulatory impact analysis for the regional NO_x SIP call. U.S. EPA Rep. EPA 452-R-98-003b. [Available from EPA/OAQPS, Research Triangle Park, NC 27711.]

——, 1999: The benefits and costs of the Clean Air Act 1990 to 2010, EPA report to Congress. U.S. EPA Rep. EPA 410-R-99-001. [Available from EPA/OAR, Washington, DC 20460.]

Ward, J. H., 1963: Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.*, **58,** 236–244.