Use of Cluster Analysis to Define Periods of Similar Meteorology and Precipitation Chemistry in Eastern North America. Part I: Transport Patterns*

MARK E. FERNAU AND PERRY J. SAMSON

Dept. of Atmospheric, Oceanic and Space Sciences, Space Physics Research Laboratory, The University of Michigan, Ann Arbor, Michigan

(Manuscript received 22 August 1989, in final form 7 February 1990)

ABSTRACT

Cluster analysis has been applied to transport vectors, derived from three years of daily backwards trajectories, in order to define a synoptic climatology of representative three-day periods of air mass movement. The resulting clusters represent groups whose mean air mass transport fields are statistically different from one another and correspond to the types of high and low pressure patterns seen on daily weather maps. Seasonal differences were evident in the frequency of occurrence of each cluster. The clusters were relatively insensitive to changes in number of sites or years used; however, different clustering methods yielded somewhat different clusters. Ward's method yielded clusters with more or less equal numbers while other methods tended to produce one large cluster and a series of "outlier" clusters. Cluster analysis was useful in the computer-assisted classification of spatial patterns of weather data and should be considered for use along with more widely used synoptic climatological tools such as principal component analysis.

1. Introduction

An aggregation approach is desirable in order to use results from the Regional Acid Deposition Model (RADM) (Chang et al. 1987) to determine seasonal or annual deposition loads in eastern North America. The complexity and number of computations involved in RADM make it costly to directly model long-term deposition by repeated simulation of episodes of several days duration (Dennis, personal communication 1988). In this paper the statistical technique of cluster analysis is applied to several years of daily transport vectors, arriving at chemistry monitoring sites in eastern North America, to identify distinct meteorological regimes and their frequencies of occurrence for future use in aggregation schemes. A companion paper will describe the precipitation patterns and chemistry associated with the resultant clusters (Fernau and Samson 1990). This process of identification of categories of atmospheric circulation type and assessment of weather and pollution elements in relation to those categories has been called synoptic climatology (Barry and Perry 1973).

Cluster analysis has been used sparingly in synoptic climatological and air pollution applications. It was used to investigate the Southern Oscillation (Wolter 1987), the dynamics of polar weather systems (LeDrew 1983, 1985), temperature trend data in the United States (Lawson et al. 1981), and the development of a long range forecast model by the British Meteorological Service (Maryon and Storey 1985). Various meteorological variables were clustered to identify areas of similar climate in Italy (Galliani and Filippini 1985) and Greece (Maheras 1984). Schulz and Samson (1988) used cluster analysis of nonprecipitating cloud frequency data to identify areas of central North America with similar cloud frequency behavior. A technique using principal component analysis to reduce the dimensionality of the data followed by cluster analysis on the important components has been used on moisture index data in India (Gadgil and Joshi 1983), precipitation data in the Mediterranean (Goossens 1985), West African climate (Anyadike 1987) and Chinese precipitation records (Ronberg and Wang 1987). Kalkstein and Corrigan (1986) presented a methodology to characterize air masses at a given locale using principal component analysis to reduce the dimensionality of surface weather variables followed by cluster analysis on the most important components. The means of the original variables and other parameters were determined for each cluster and a physical explanation was attached to the results. Kalkstein et al. (1987) refined the method and demonstrated it for two other sites. Moody and Samson (1989) applied cluster analysis to stratify meteorologically similar events. For one site they used 850 hectopascal (hPa) temperature, wind vectors and mixing ratio to separate different air masses. At two other sites, air mass tra-

* This work was not funded through Argonne National Laboratory.

Corresponding author address: Dr. Mark E. Fernau, Environmental Assessment and Information Sciences Division, 362, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439.

jectories were clustered into events with similar transport history. The area of origin and speed of movement of the trajectories differed for each cluster. Cluster analysis has also been used to examine air pollution variables, alone and with the associated meteorology (see Fernau and Samson 1990, for a review).

The work described in this paper demonstrates that cluster analysis is a viable technique for classifying weather events into distinct groups whose members exhibit similar spatial patterns of transport variables and explores the sensitivity of the cluster analysis to methods used, variables used, and other parameters that must be subjectively specified. As demonstrated by Moody and Samson (1989) and Kalkstein et al. (1987) for single sites, cluster analysis shows promise for achieving the goal of stratifying events (cases) by meteorological parameters. In this paper its application to a larger spatial domain is investigated and the transport characteristics of the clusters resulting from various analyses are described. The precipitation and chemical characteristics are described in Fernau and Samson (1990). A more fully detailed and illustrated treatment of the clustering work can be found in Fernau (1988).

2. Data and methods

a. Upper air data and trajectories

Upper air data tapes (TD-973, BACKWARD NAMER-WINDTEMP) containing processed rawinsonde and pibal observations extracted from the raw soundings for all stations located in North America (Alaska excluded) were obtained from the NOAA National Climatic Center in Asheville, North Carolina for the years 1979–83. These data were used as input to a version of the Atmospheric Resources Laboratories Atmospheric Transport and Dispersion Model (ARL/ATAD) trajectory model (Heffter 1980) to calculate back trajectories. The model calculates trajectories arriving at a receptor every six hours (0000, 0600, 1200, 1800 UTC). For each trajectory the location is calculated every three hours upwind, backwards in time to a maximum of 72 hours. The travel for each three-hour segment is determined using the upper air information from each rawinsonde station within a 250 nautical mile radius of the previous end point, using a r^{-2} weighting where r is the distance of the station from the end point. Linear interpolation in time is also done because the data are only available every 12 hours. If there are not sufficient data within the specified radius or if the trajectory travels to a point outside of the grid domain the trajectory terminates prior to reaching 72 hours upwind. The model only considers upper air data located in and averaged through the "mixed layer," defined as extending from the surface to the first non-surface-based inversion (Heffter 1980). If no inversion exists the layer top is set at 300 m above ground level. Thus each trajectory is steered by the average winds in the mixed layer.

Trajectories were calculated for 1979 through 1983 at 22 sites for which precipitation chemistry measurements were available (Fig. 1). The trajectory files consisted of one file per site per year from 1979 to 1983. Each file had one trajectory per day arriving at 1800 UTC (1300 EST), represented by latitudes and longitudes or missing value indicators for each three-hour interval out to three days upwind. For the purposes of this work, the interval latitudes and longitudes were converted to straight-line distance vectors from the receptor origin to each interval end point. For specific experiments, the distance vectors for a specific upwind

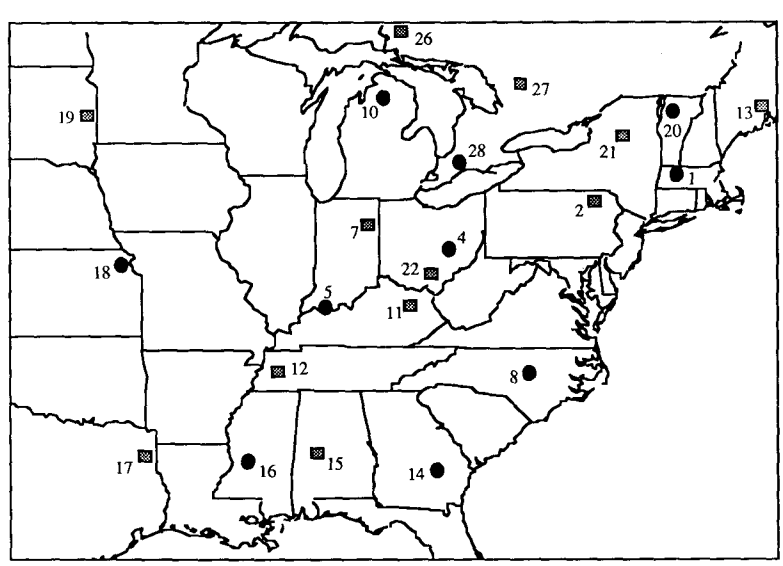

FIG. 1. Locations of trajectory receptor sites. Sites with circles are the subset denoted as group A in the text.

interval for each site were merged into one file, arranged in chronological order. The data were arranged in moving three-day period format, six variables per record (y and x components for days $n-1$, n, and $n+1$), for subsequent cluster analysis studies. The three-day periods allow compatibility with RADM, which is generally exercised to model three-day periods, and also allow the movement of the systems to be seen. Because of computer memory constraints, it was necessary to use limited subsets of this trajectory data for any given experiment.

b. Cluster analysis

Cluster or linkage analysis is a collection of statistical methods whose purpose is to divide a dataset into groups of similar cases or variables, hopefully reflecting underlying structure present in the data. The goal is to have group members differ as little as possible and to have each group be as distinct from the other groups as possible. In implementing cluster analysis one must choose a method of determining similarity among clusters and a method of determining the distance between pairs of points. Distances between points are usually determined using straight line or Euclidean distance measures or correlation coefficients. Common clustering methods to determine between-group distances include single, complete, and average linkage methods, centroid and median methods, and Ward's or minimum variance methods.

The major clustering method used in this work is Ward's method, although several other methods were tested in sensitivity studies. Ward's method always uses Euclidean distance. It joins clusters whose merger results in the minimum increase in the sum over all clusters of the within-group sum of squares. The within-group sum of squares, also known as the cluster variance or within-group sum of error squares, is the sum of squared Euclidean distances from the data points to the mean vector of the cluster (Ward 1963). Details of Ward's and the other methods used as well as discussions of the myriad of other available techniques can be found in Anderberg (1973) and Gordon (1981).

Cluster analysis, although computer-based, involves some subjective decisions regarding the choice of variables to cluster, which type of cluster analysis to use (agglomerative vs. divisive; hierarchical vs. iterative; defined below), which method and distance measure to use, and at which step to terminate the clustering process, i.e. determining the optimum number of clusters for the dataset. Once variables, type of clustering and methods have been selected, the major difficulty in reproducing a result would be in following the same procedure for termination of clustering. If differing types or methods give differing results one must use knowledge of the problem to determine which is more physically realistic.

All cluster analysis in this study was performed using the University of Michigan Statistical Research Laboratory Michigan Interactive Data Analysis System (MIDAS) (Fox and Guire 1976). The merged distance vector file was read into MIDAS, cluster analysis was performed, and the group memberships of each event at the last fifty steps were saved for use in further analysis of the resulting clusters. Clustering was done in a casewise manner (as opposed to clustering by variables), with a case being defined as a three-day period or event. All clustering in this study was hierarchical (clusters, once merged, cannot be separated) and agglomerative (going from many clusters to one large cluster). The distance measure used was Euclidean distance, except that the correlation coefficient was used in a sensitivity study. The variables were all weighted evenly and were not standardized. Standardization (conversion to mean of zero, standard deviation of one) is generally recommended if the variables being clustered are of varying units or orders of magnitude.

Part of the output from MIDAS is the distance measure value used at each step to join clusters. When plotted versus step number, this information can give guidance as to the optimal number of clusters. Sudden increases in the value of the distance measure reflect the joining of clusters which are not very similar.

3. Choice of clustering data

An initial attempt to cluster on precipitation totals at various sites or grid cells gave discouraging results, with clusters representing periods of uniformly high precipitation over the entire domain or else high precipitation at a single grid cell or station with little precipitation elsewhere. Efforts were then concentrated on classifying the transport patterns over eastern North America by using distance vectors arriving at the 22 sites. A characteristic of cluster analysis is that for a record to be included in the analysis it must have no missing variables. The number of records available for clustering was maximized by eliminating sites with large amounts of missing data and examining only the first few upwind time steps. However, there is also a limit on the allowable size of the data matrix used in the clustering routine. After some experimentation a balance was reached among site number, upwind distance, and number of years included so as to maximize the spatial coverage and duration of the distance vectors and the time period covered while still keeping the problem tractable.

A subset of ten sites, chosen so as to maintain as well as possible the original spatial domain (Fig. 1), was selected and three years of data were used. The years 1979, 1981 and 1983 (a total of 1093 three-day moving periods) were selected because most existing RADM simulations were contained therein. A case or record consisted of six variables for each site: the x and y components of the distance vector originating 12

hours upwind and arriving at the site at 1800 UTC for each of the three days comprising an event—a total of 60 variables per event. Twelve hours travel time was found to be sufficient to allow the weather patterns to be seen. The above combination of parameters led to a data loss of only 27%. Other combinations of number of sites, years used and upwind distance were examined in sensitivity analyses.

4. Results

When discussing clusters, step numbers are defined by the number of clusters existing at that point. The number labels assigned to the clusters are arbitrary. Using the MIDAS output of total within-group error sum of squares at each step number, likely breakpoints for terminating the clustering operation were selected. Based on the two factors of sharp changes in slope and retention of a workable number of clusters, steps 7 and 18 were chosen for further examination.

a. Analysis of step 7

1) MEAN TRANSPORT

The step 7 clusters can be characterized by plots of the mean transport vectors for each cluster. All sites were plotted although the clusters were determined only by ten sites. In the following figures, the length of each arrow is the mean distance traveled during the 12 hours prior to arrival at the receptor. The arrow heads are located at the receptor and are proportional to travel distance. The numbers at the ends of the arrows indicate to which day of the three-day period the map corresponds. Figures of median flow are not shown but are nearly identical.

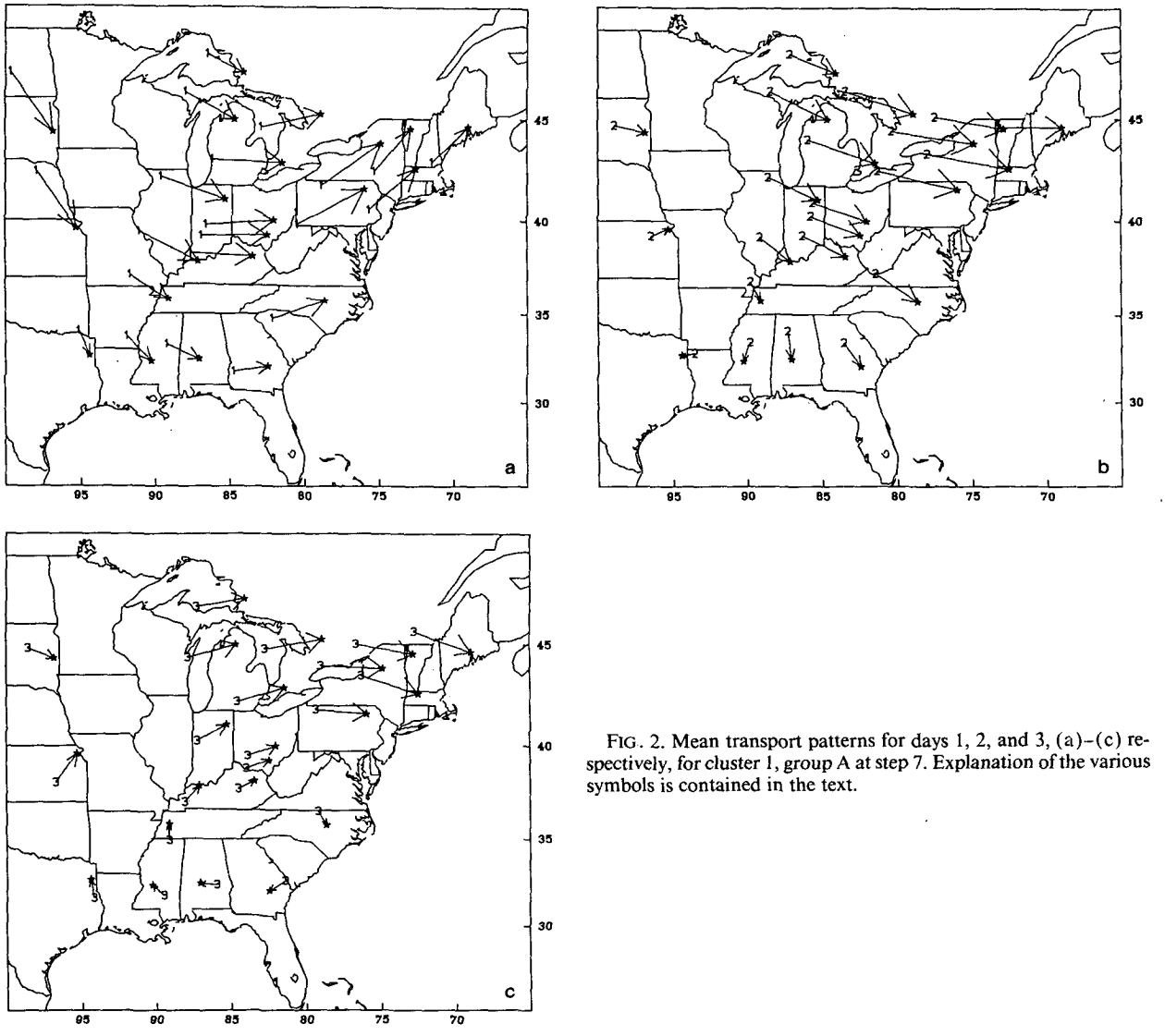

FIG. 2. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 1, group A at step 7. Explanation of the various symbols is contained in the text.

The patterns seen in these figures resemble weather systems occurring in the atmosphere. Figures 2a–2c depict the flow for cluster 1. The period begins with a trough extending in a north–south direction from Michigan to Alabama. The trough moves eastward off the coast by day 2 with a center of high pressure located in the Arkansas area. By day 3 the trough leaves the domain and the high pressure becomes well established in the southeast. Air movement is quite rapid around the trough, especially behind it, but slows by day 3 in the area around the high.

Cluster 2 (Figs. 3a–3c) is similar to cluster 1. On day 1 a trough with rapid air movement extends north–south from Minnesota to Louisiana. Days 2 and 3 are very similar in appearance to days 1 and 2 of cluster 1, with the trough moving east and then a high developing in the south. It seems likely that cluster 1 events frequently follow cluster 2 events. The transition matrix will be examined below.

The transport of cluster 3 is shown in Figs. 4a–4c. Day 1 has fairly weak clockwise flow around a closed high located in the southeast. During the next two days the high slowly moves east to a position over the ocean just off the Carolinas. Somewhat stronger southwesterly flow on the back side of the high is dominant and there is a hint of a trough in the Minnesota area.

The transport of cluster 4 is represented in Figs. 5a–5c. In the mean, cluster 4 comprises elements of both clusters 3 and 2. Day 1 has southwesterly flow behind a high off the Carolinas with a trough in the Minnesota region. On days 2 and 3 the north-south oriented trough moves east into Ontario, extends to the Gulf coast, and wind speeds become more rapid. Clusters 3 and 4 tend to have large amounts of pollutant deposition associated with them (Fernau and Samson 1990).

Figures 6a–6c have the flow for cluster 5. New England and Ontario are characterized by moderate northwest flow, slowing by the third day. The rest of

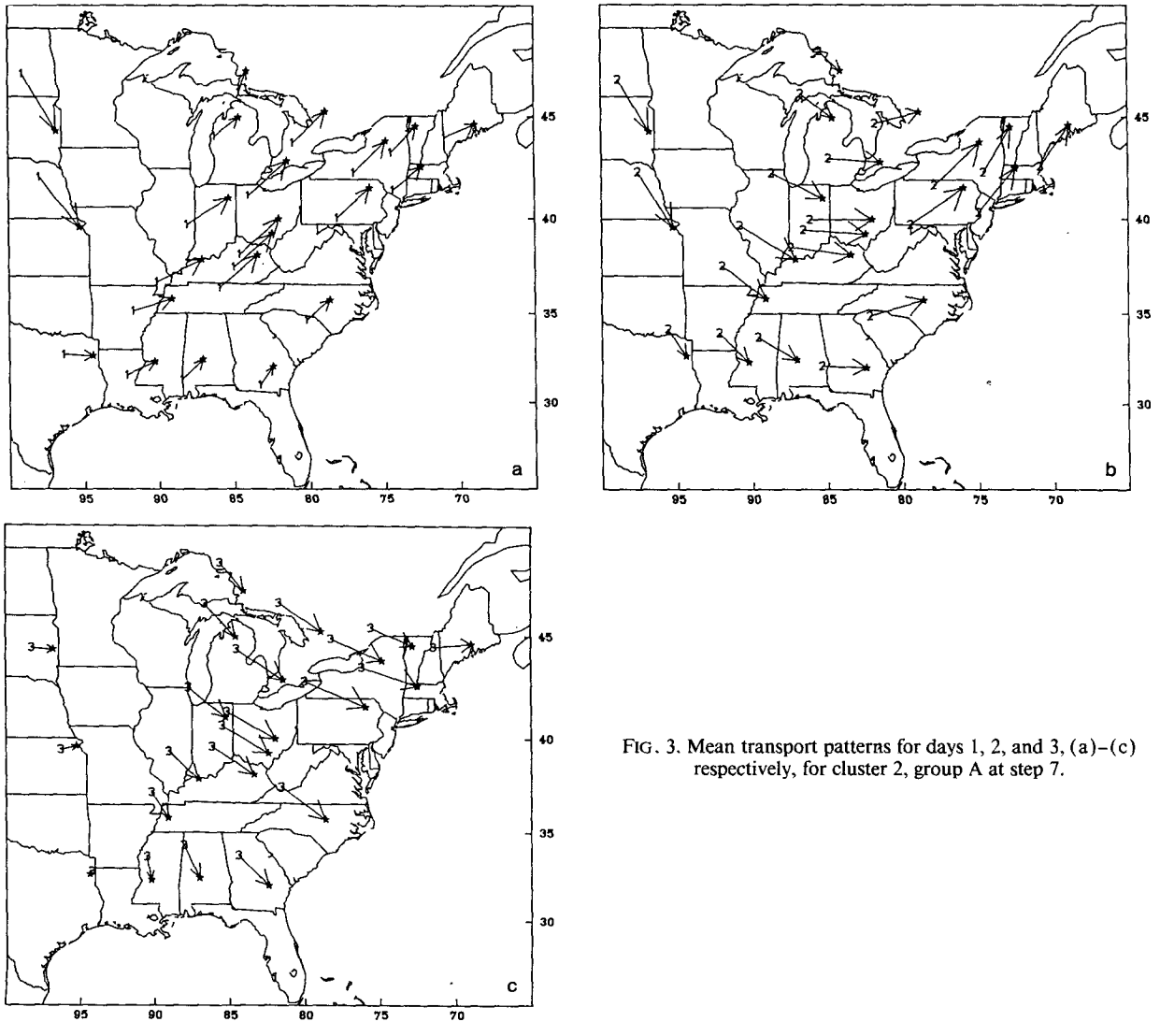

FIG. 3. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 2, group A at step 7.

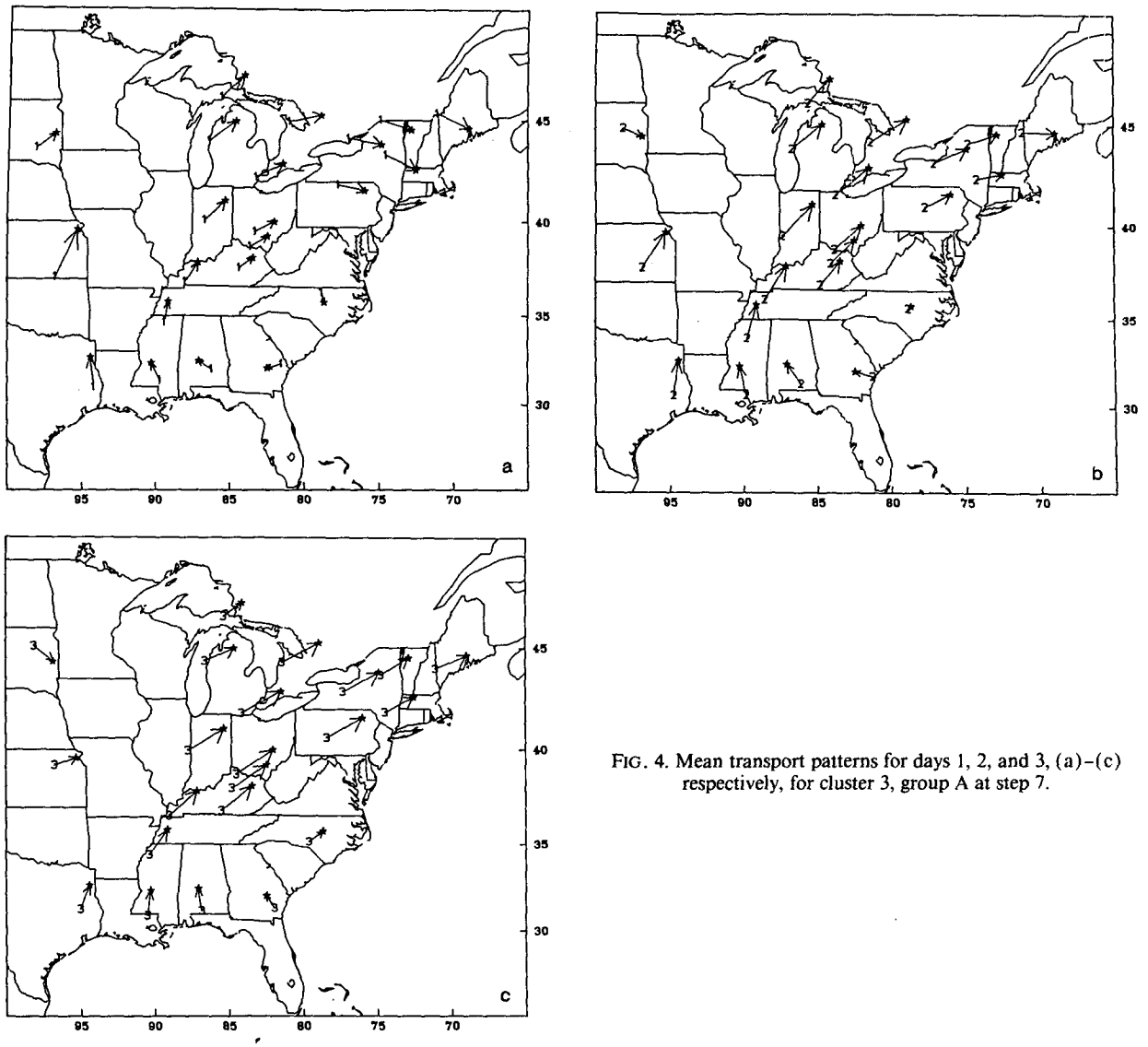

FIG. 4. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 3, group A at step 7.

the region is dominated by stagnant anticyclonic flow as a high pressure center slowly drifts eastward from Illinois to New York during the period.

Cluster 6 begins with a low pressure center located in New England and moderate northwest flow over the remainder of the region (Figs. 7a–7c). By days 2 and 3 the low has moved off the map, winds have a more northerly component, and a high pressure center begins to move from Texas into Louisiana. One would expect this cluster to represent clean continental polar air masses intruding into the United States.

The flow pattern for cluster 7 (Figs. 8a–8c) remains very similar over all three days. Winds diminish and become relatively stagnant and ill-defined by the third day. There is a ridge present in the center of the region.

These figures show that cluster analysis of distance vectors has classified events into distinct and realistic meteorological categories, at least as depicted by the mean transport. Figure 9 shows an example of the standard deviations associated with the mean transport vectors for cluster 1. Crosses reflecting the magnitudes of the standard deviations in the x and y directions are located at the end of each arrow. It is clear that the variation around the mean is substantial, although the sense of the flow (cyclonic vs anticyclonic) is preserved in most cases. The within-group standard deviation averaged by cluster at each site and then again over the 22 sites was 188 km in the y direction and 178 km in the x direction at step 7. Among-group standard deviation averaged by site was 135 km in the y direction and 104 km in the x direction. Within-group standard deviation was larger than among-group standard deviation, indicating a poor degree of separation among clusters.

A multivariate one-way analysis of variance was used to test whether mean vectors were the same in all clus-

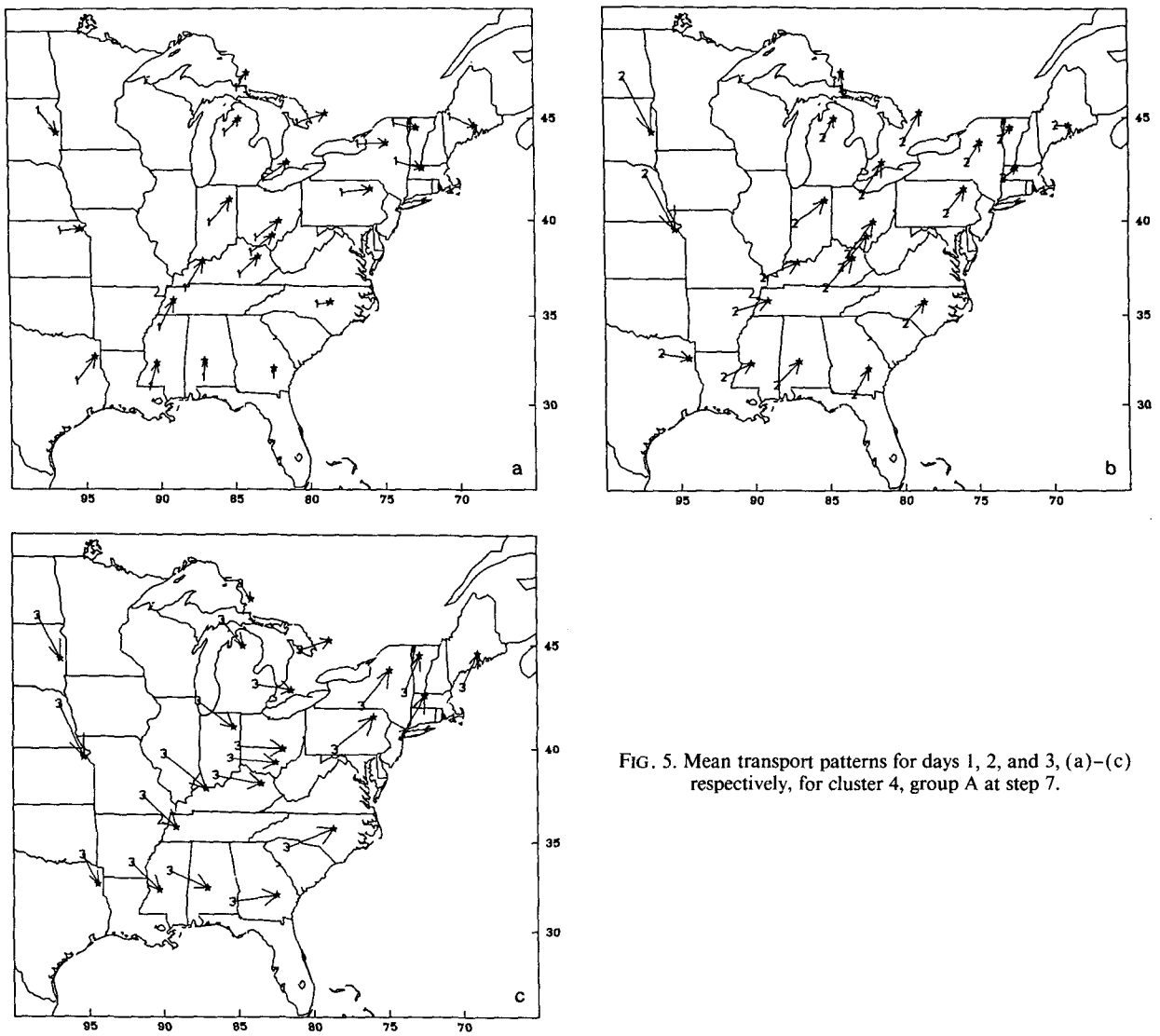

FIG. 5. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 4, group A at step 7.

ters, a vector in this case meaning a series of variables as defined below. If there were significant differences among clusters then Hotelling's T-square test was used to perform a multivariate two-sample comparison of whether the mean vectors of specific pairs of clusters were different. Both tests were applied to one site at a time with the mean vectors for each cluster consisting of the means from the six distance components of the three days comprising an event. The results of the tests were that for all twenty sites tested, both the ten defining the cluster and the ten others, there were significant differences in the mean vector among clusters at the .99 confidence level (α = .01). Individual T-square tests showed that in almost all possible cases pairs of clusters had different mean vectors at the .99 level. Although the cluster distributions overlap somewhat and have large standard deviations, they are statistically different in their transport patterns. It should be noted that there is uncertainty in the trajectories themselves due to the low resolution of the upper air network. The horizontal error in trajectory location can be 100 to 200 km 24 hours upwind and as large as 100 to 500 km at 72 hours upwind, depending on the meteorological conditions (Kahl and Samson 1988).

2) DESCRIPTIVE STATISTICS FOR STEP 7

The distribution of cluster membership by month and year is given in Table 1 for step 7. Percentages for nonmissing clusters are calculated on the total, nonmissing plus missing. Sixty percent of the eliminated records come from 1979, with almost half of the possible 1979 three-day periods eliminated. The data retention record for the other years, 80 percent or better, is far superior. The majority of the missing records are from the colder months when wind speeds are generally

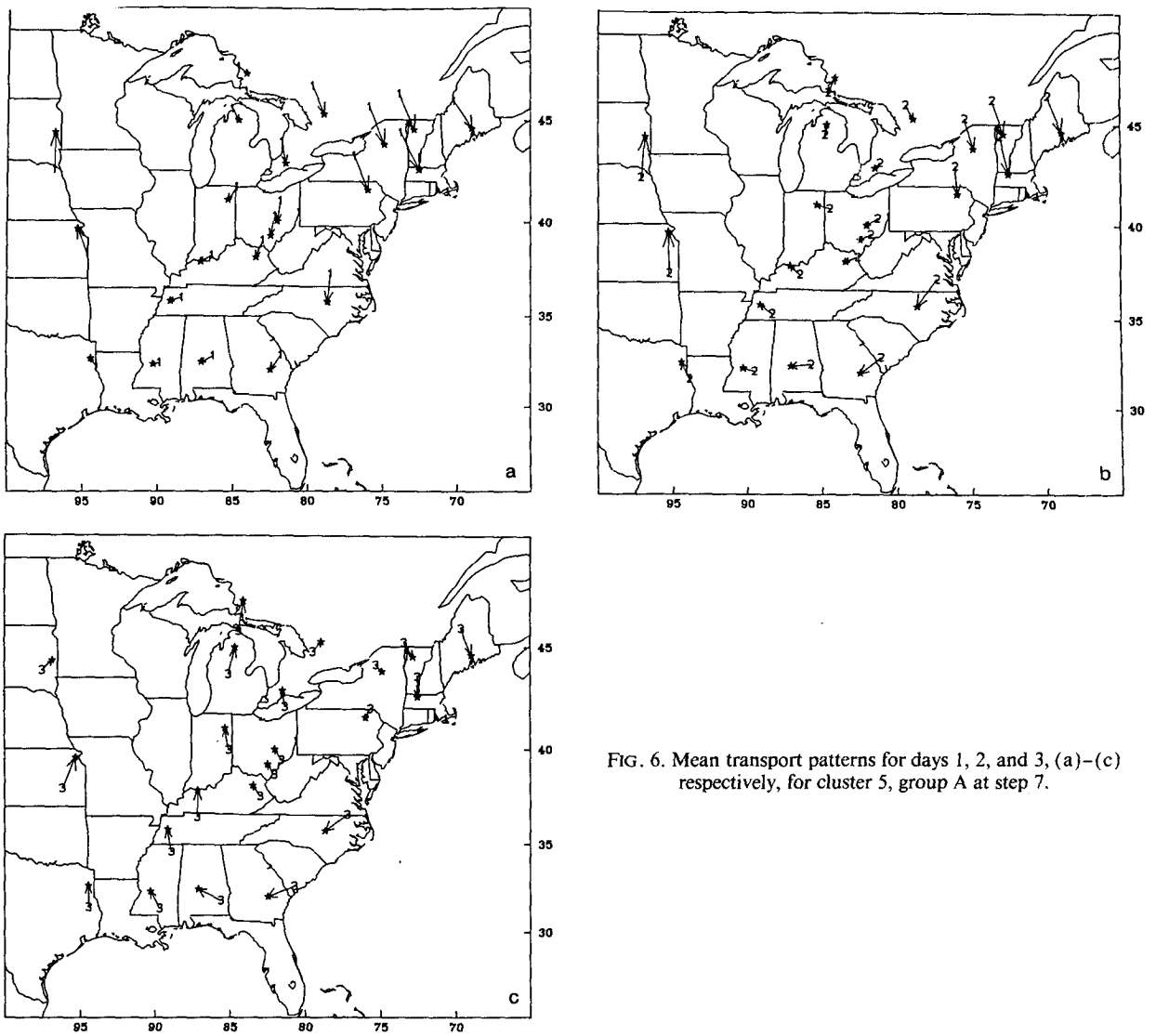

FIG. 6. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 5, group A at step 7.

greater and the trajectories are more likely to exit the model grid. Year-to-year variability in the overall cluster memberships is not very large and seems to be mostly due to the missing 1979 data, as all seven clusters have the least number of members in that year. Except for cluster 2, the cluster membership does not change much from 1981 to 1983. Accounting for the variation in missing data, the clusters dominated by anticyclonic flow (3 and 5) seem more common in 1983 than in 1981. The two most common clusters are 3 and 7, both of which have westerly components. It is possible that cluster 7 is representative of average conditions during the year or periods when there are no strongly defined flow features. Cluster 5, the stagnant high, occurs about 10 percent of the time overall, and the cyclonic clusters occur less frequently.

Regarding seasonal variability, clusters 1 and 2 rarely occur in the warmer months. Clusters 4 and 6 also are less likely to occur in the warmer months. Cluster 6 is most commonly found from January through April. These results reflect the rareness of organized cyclonic storms in the warm months. Clusters 3 and 7 are common all year but have maximum occurrences in the warm months. Cluster 5, the slowly moving high, is more common in the warm months. Clustering of transport vectors was never stratified by season in this study. Since the vectors encompass both direction and speed it was hoped that seasonal differences in transport would be reflected in the clusters themselves without explicitly being accounted for. This seems to be the case and is more obvious at higher steps.

Table 2 gives the cluster transitions from one period to the next. From this table and the mean plots one can see how the clusters overlap and interact with one another. Some of the similarities noted in the figures are confirmed. Cluster 2 is most often followed by cluster 1, as suspected from the mean spatial patterns. Similarly, cluster 3 does tend to change to cluster 4.

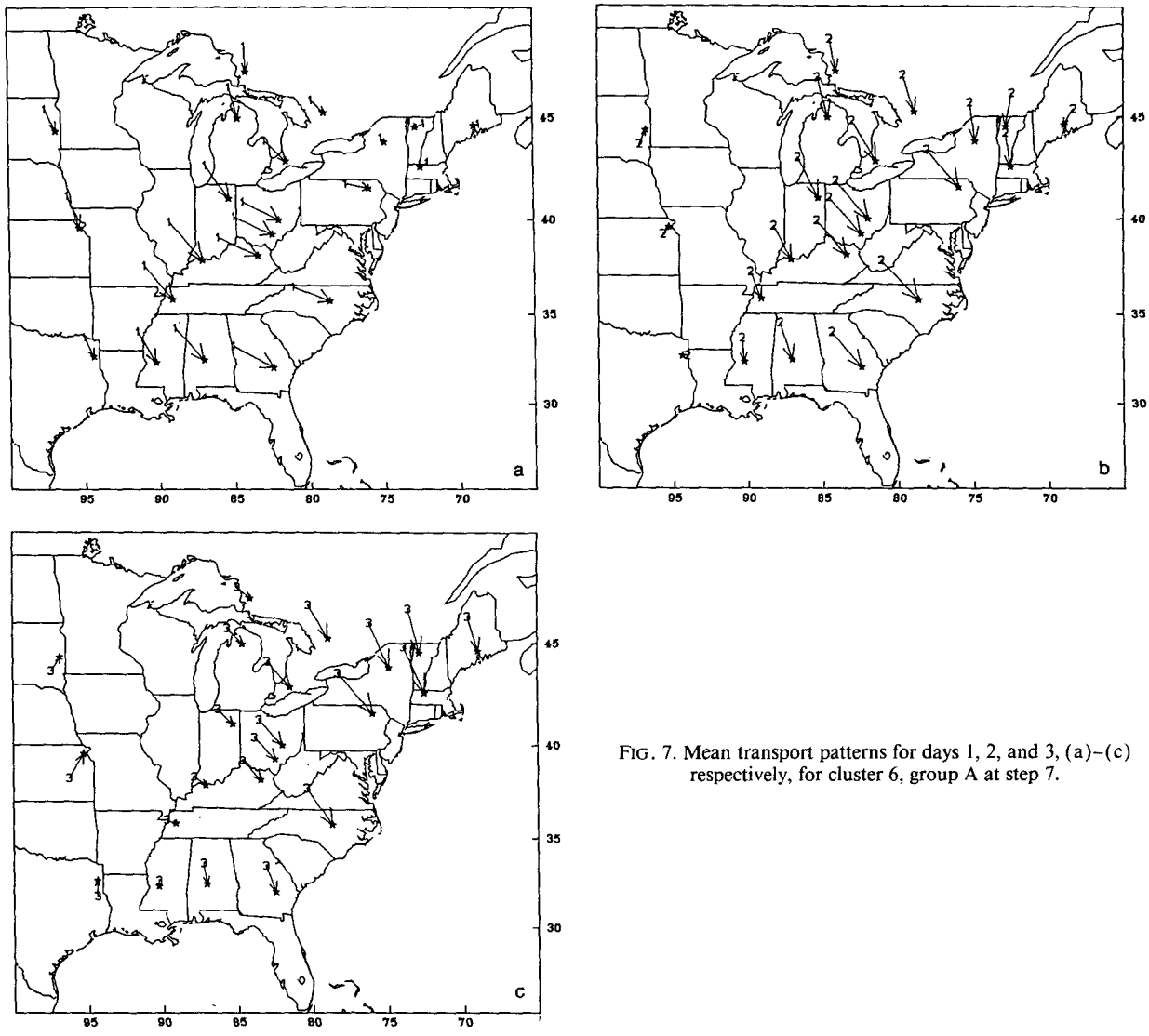

FIG. 7. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 6, group A at step 7.

Cluster 2 never changes directly to the anticyclonic clusters 3 and 5. Cluster 4 is never followed by the high pressure of cluster 5. Cluster 5, in turn, is never followed by the cyclonic flow of clusters 1, 2, or 6. As one might expect, the high pressure of clusters 3 and 5 and the stagnant, unorganized flow of cluster 7 tend to persist with all three clusters being most likely to follow themselves at least once. It is clear why clusters 3 and 4 are prime candidates to be polluted air masses because stagnant clusters 5 and 7 tend to change to cluster 3, cluster 3 is quite persistent itself, and then it is likely to change to cluster 4, which is associated with large precipitation amounts (Fernau and Samson 1990).

3) EXPLORATION OF THE MISSING DATA

It was shown earlier that the missing periods made up about 30% of the total periods, varying by year and month. Examination of the mean transport plots shows that eastern coastal lows may not be well represented. The missing data were examined more closely to see if they represented distinct meteorological groups. Figures 10a–10c show the mean flow at 12 hours upwind for the periods of missing data excluded from cluster analysis. That is, they represent the flow at the nonmissing sites when one or more sites had missing data. The pattern does not change from day 1 to day 3 and seems to represent the average westerly flow found during the course of a year. It can not be determined from the figures whether systematic transport biases exist in the missing data. Missing easterly flow would have the effect of rotating all mean vectors to the west; this could be the case in the figures. The standard deviations are large compared to the other clusters (over 200 km). An inspection of National Weather Service Daily Weather Maps confirmed that many coastal storms were missing. Those not missing tended to be assigned to cluster 6.

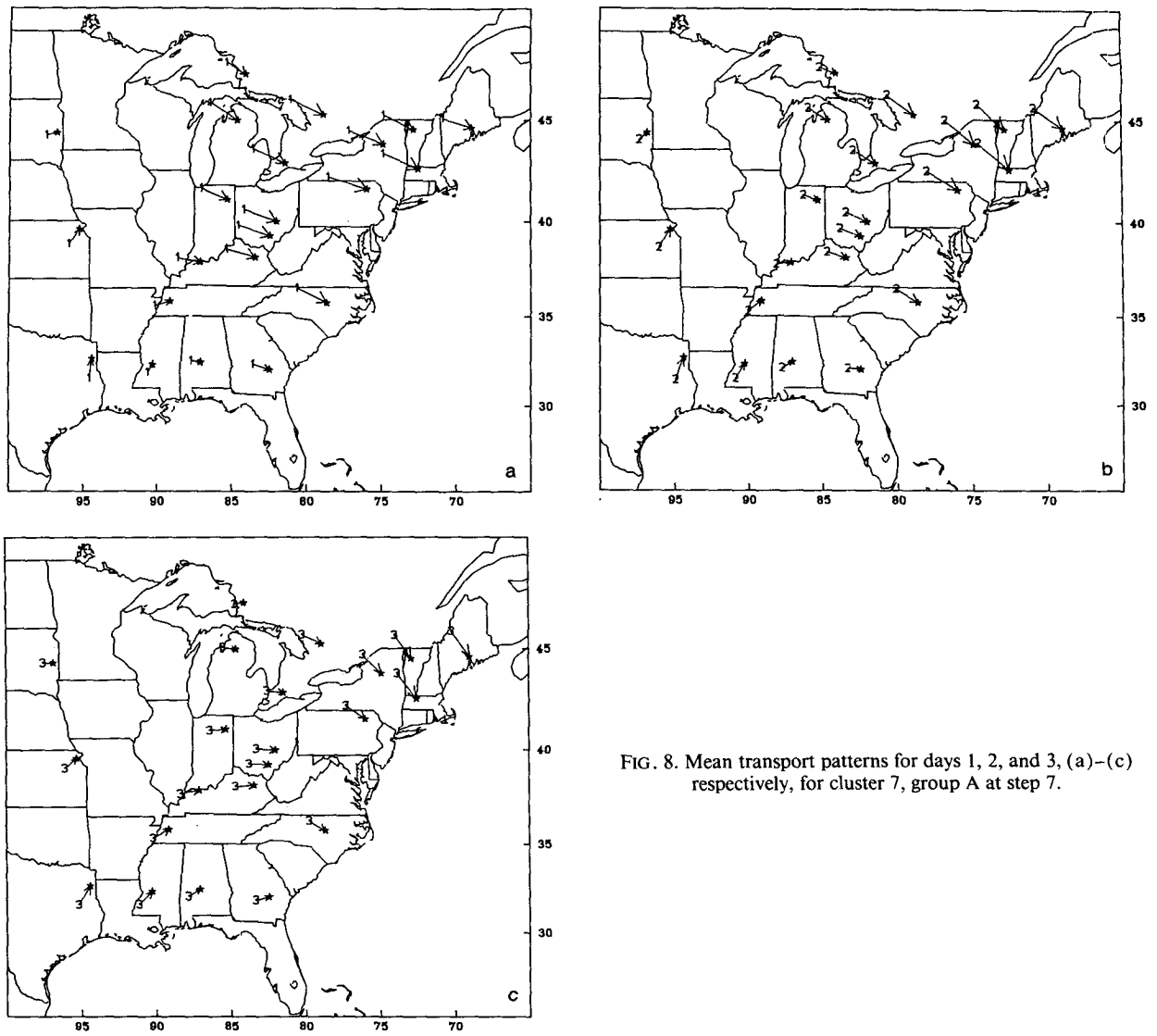

FIG. 8. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 7, group A at step 7.

b. Analysis of step 18

Based on the change in the MIDAS output of total within-group error sum of squares at each step number, step 18 was also a reasonable cutoff for terminating the clustering operation. The clusters at 18 were compared to step 7.

1) MEAN TRANSPORT

The transport patterns at step 18 can be compared to step 7. As with step 7 clusters, almost all possible combinations of clusters are significantly different from each other at the .99 confidence level, using the Hotelling's T-square test. Examination of the Step 18 mean transport patterns illustrates both the ability of Ward's method to distinguish between similar patterns with wind speed differences and the fact that the step 7 clusters obscure within them some interesting differences in flow and likelihood of persistence which become evident at higher steps. In using cluster analysis one must strike a balance between trying to derive a few basic flow regimes which represent the important features and losing subtle but possibly important features.

Cluster 1 of 18 is identical to cluster 1 of 7. Cluster 2 of 7 is composed of clusters 2, 3, and 4 at step 18. They all retain the north–south trough moving across the region but have differences in wind direction and speed, speed and direction of system movement, position of the low, and likelihood of persisting more than one period. Cluster 3 at step 7 combines five step 18 clusters, all of which hold in common the high pressure center located in the southeastern United States. However, the wind speeds and exact position and subsequent movement of the high differ in each cluster and three of the clusters are three to five times more likely to persist than the other two. In cluster 9 the high is

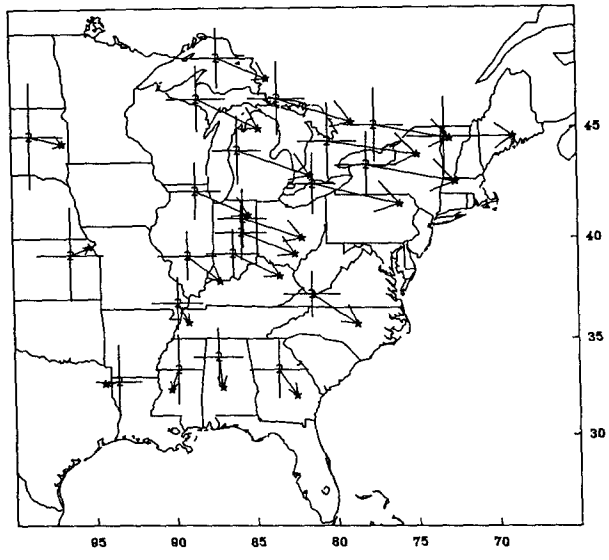

FIG. 9. Standard deviations associated with group A, step 7 Ward's method in the x and y directions at each site for day 2, cluster 1.

initially located in West Virginia and remains almost stationary, perhaps even moving a little south and west. Winds are sluggish through the period and the northwest trough never appears (Figs. 11a–11c).

Clusters 10 and 11 both resemble cluster 4 at step 7 in pattern, but cluster 11 has much more rapid wind speeds and never occurs consecutively. Clusters 12 and 13, which comprise cluster 5 of 7, show that cluster's slowly eastward drifting high pressure center in the north. The chief differences between the two clusters are that cluster 12 seems to represent the same system lagged one day behind cluster 13 and cluster 12 is only half as likely to exist for more than one consecutive period. Clusters 14 and 15 at step 18 combine to form cluster 6 of 7. Cluster 14 begins, in the mean, with a low located in New England. It moves east, northerly flow occurs over most of the map region, and finally a ridge is seen in the Mississippi valley. Cluster 15 has a closed cyclone over Lake Erie which moves east off the New England coast, leaving northerly flow over most of the eastern United States. This pattern is ob-

TABLE 1. Step 7 cluster membership by month and year. Percentages are for each year and total (Σ).
M: Missing periods, not assigned to a cluster.

Cluster	J	F	M	A	M	J	J	A	S	O	N	D	Number	Percent	Year
1	5	0	1	0	0	0	0	0	1	1	5	2	15	4.1	1979
	0	4	2	5	0	2	0	0	2	3	1	2	21	5.8	1981
	4	0	0	1	2	2	0	0	3	3	0	4	19	5.2	1983
	9	4	3	6	2	4	0	0	6	7	6	8	55	5.0	Σ
2	2	0	3	3	1	2	0	0	1	3	4	3	22	6.0	79
	3	5	3	6	4	3	1	1	3	4	3	5	41	11.2	81
	4	1	1	4	3	0	2	0	2	1	0	4	22	6.0	83
	9	6	7	13	8	5	3	1	6	8	7	12	85	7.8	Σ
3	2	0	0	4	4	10	8	11	1	6	7	4	57	15.7	79
	2	6	5	8	3	12	4	9	4	7	4	3	67	18.4	81
	3	8	2	6	7	13	10	8	11	6	0	2	76	20.9	83
	7	14	7	18	14	35	22	28	16	19	11	9	200	18.3	Σ
4	4	1	2	4	2	0	1	0	1	3	2	2	22	6.0	79
	3	6	3	3	5	2	2	5	2	4	1	4	40	11.0	81
	5	1	4	4	5	3	0	0	1	3	1	5	32	8.8	83
	12	8	9	11	12	5	3	5	4	10	4	11	94	8.6	Σ
5	2	1	1	1	0	4	3	3	0	1	0	1	17	4.7	79
	0	0	0	1	8	2	5	9	2	6	6	0	39	10.7	81
	3	1	6	1	4	8	5	5	6	8	0	3	50	13.7	83
	5	2	7	3	12	14	13	17	8	15	6	4	106	9.7	Σ
6	3	3	1	4	0	2	0	0	0	0	0	0	13	3.6	79
	8	3	6	1	2	1	1	0	4	3	2	4	35	9.6	81
	3	2	6	6	1	0	2	2	0	2	0	1	25	6.9	83
	14	8	13	11	3	3	3	2	4	5	2	5	73	6.7	Σ
7	1	4	1	1	5	3	14	6	1	0	2	3	41	11.3	79
	8	1	5	6	5	8	15	7	9	1	7	3	75	20.5	81
	2	5	5	5	3	4	12	16	7	1	1	5	66	18.1	83
	11	10	11	12	13	15	41	29	17	2	10	11	182	16.7	Σ
M	11	19	22	13	19	9	5	11	25	17	10	16	177	48.6	79
	7	3	7	0	4	0	3	0	4	3	6	10	47	12.9	81
	7	10	7	3	6	0	0	0	0	7	28	6	74	20.3	83
	25	32	36	16	29	9	8	11	29	27	44	32	298	27.3	Σ

TABLE 2. Number of direct transitions from cluster to cluster at step 7.

	To						
From	1	2	3	4	5	6	7
1	9	2	19	3	3	1	12
2	33	18	0	2	0	19	6
3	2	13	97	38	9	3	24
4	6	43	2	20	0	7	10
5	0	0	35	4	55	0	4
6	2	0	1	2	16	29	19
7	0	3	37	18	16	9	87

scured at step 7. Clusters 16, 17 and 18 combine to form cluster 7 at step 7. The pattern of cluster 7, step 7, seems to be an average of the three clusters at step 18 rather than a distinct pattern. Clusters 16, 17, and 18 all have different median patterns, with 16 showing higher wind speeds and a transition from northerly to westerly flow, 17 being extremely stagnant over the entire region with little change in a weak ridge pattern, and 18 showing weak westerly flow. Cluster 16 repeats only a quarter to a third as often as the other two.

c. *Comparison of mean transport patterns to daily weather maps*

The mean transport patterns can be compared to National Weather Service Daily Weather Maps to see if they reflect the observed pressure fields. At step 18, the daily maps resemble the mean patterns in most cases examined. The resemblance can frequently be seen in the surface pressure pattern and wind field. Other times the cluster means resemble more the upper air height pattern. This is a reflection of the fact that the trajectories are calculated from winds averaged through the mixed layer. The daily maps portray the

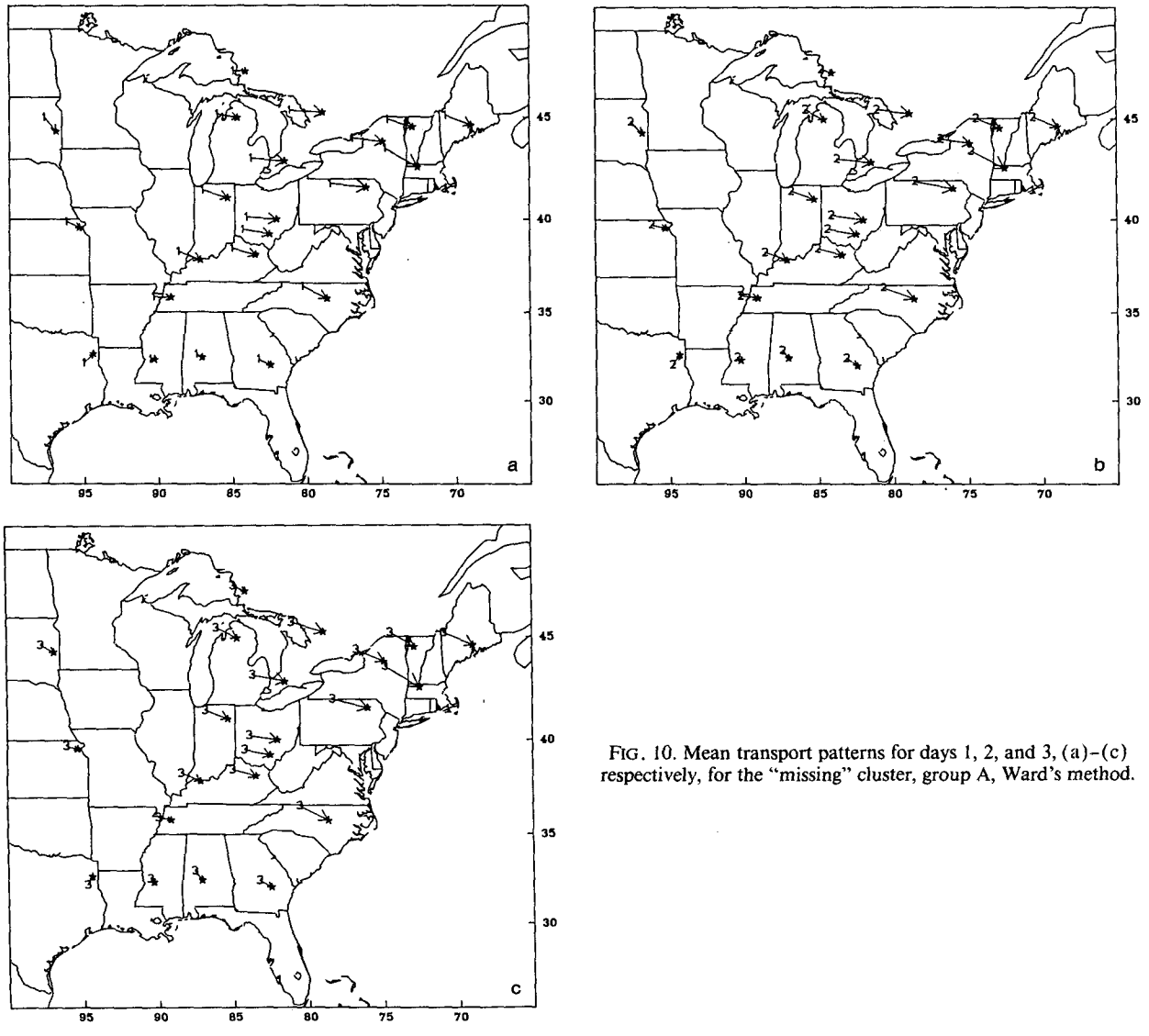

FIG. 10. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for the "missing" cluster, group A, Ward's method.

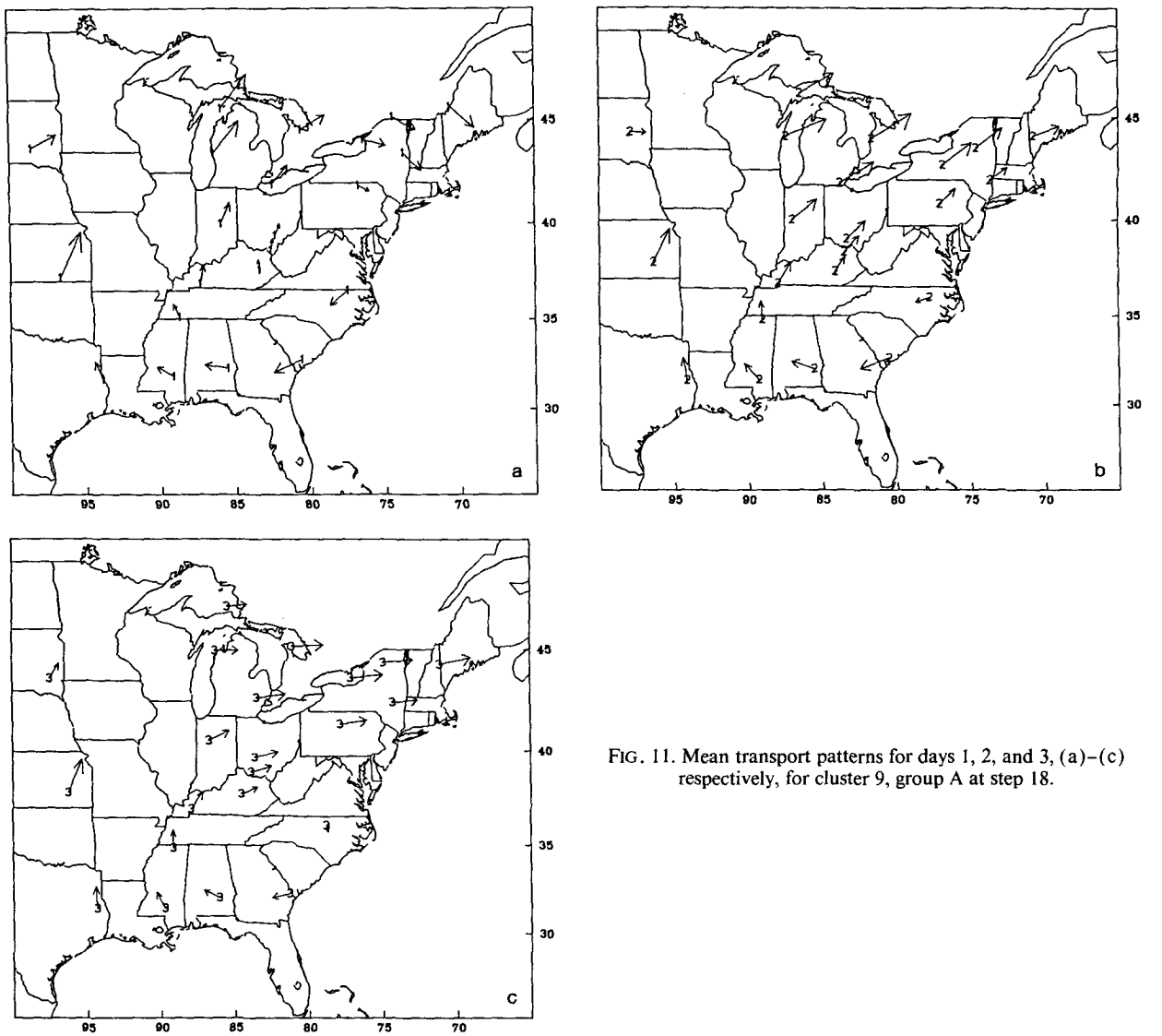

FIG. 11. Mean transport patterns for days 1, 2, and 3, (a)–(c) respectively, for cluster 9, group A at step 18.

meteorology as it existed at 1200 UTC so they do not correspond exactly in time to the mean plots.

As one might expect, there is less agreement between cluster means and maps at step 7, especially for clusters such as 3 and 7 which are composed of several step 18 clusters with differing flow patterns. However, the locations of pressure features and relative wind speed and direction on many surface and upper air maps still correspond well to the mean patterns. Cluster 1, which does not change from step 18 to step 7, often agrees with the maps, with many of the maps showing a north–south oriented cold front moving through the Appalachian region and then being replaced by high pressure and low wind speeds in the lower Mississippi Valley and the southeast states. The transitory nature of clusters such as step 7 cluster 1 and several of the step 18 clusters is supported by very strong upper-level winds and surface lows moving rapidly across the map region on the corresponding maps.

An example where the observed patterns correspond to the mean cluster pattern is now described in more detail. Nine consecutive moving three-day periods were assigned to cluster 9 at step 18 (Figs. 11a–11c), extending from 9 to 17 June 1983. Cluster 9 is seen from the transition matrix to be very persistent. It is quite enhanced in sulfate deposition relative to precipitation and also precedes clusters 5, 12 and 18, which are themselves enhanced (Fernau and Samson 1990). From the observed flow it is easy to see why this is so. The 500 hPa flow during the eleven days comprising the cluster 9 period is unorganized with extremely light wind speeds in the eastern United States. For much of the period there is a huge ridge over the area and the polar front in eastern North America is located well into Canada. The ridge is stationary until the final three days of the period when wind speeds increase slightly and the ridge begins moving eastward. At the surface, in the area east of the Mississippi River, winds are light,

precipitation is nonexistent, skies are clear at many stations and high pressure dominates the entire time, with stationary fronts located in part of that area on some days. The center of the high is located over the Mid-Atlantic states and shows little or no movement for most of the period, just as portrayed in the mean pattern. Figure 12 shows the surface pressure field partway through the episode. Finally, by the end of the period a cold front extending from Canada moves east across the region and triggers showers and thunderstorms. Haze and fog begin to be reported on 10 June and quickly spread eastward and northward over wide areas of the eastern United States and Canada, eventually being swept east ahead of the cold front. These conditions all appear to be very conducive to or indicative of a regional buildup of sulfate. The chemistry of this cluster is discussed in more detail in Fernau and Samson (1990); here it is sufficient to note the correspondence between the actual weather pattern and the cluster characteristics.

d. Sensitivity of cluster analysis to various parameters

One of the goals of this work was to test the sensitivity of the cluster analysis results to changes in factors such as method of clustering or distance measure, number of sites used and time period. This section details the results of some sensitivity experiments.

1) 18-HOUR UPWIND CLUSTERS

The clustering in the previous section involved trajectories originating 12 hours upwind. The step 7 results from that clustering will be denoted in the following section as S7. Three-day clustering was also done using Ward's method on vectors originating 18 hours upwind for all 22 sites for 1979–1983. At step 6 of this clustering run mean patterns similar to S7 were found. Cluster 1 is the high pressure in the southeast, southwest flow around it, and a trough in the northwest as the event evolves. Cluster 2 depicts a central trough. Cluster 3 is analogous to cluster 4 of S7 with the southeast high moving east and being followed by an eastward moving trough. Cluster 4 is the eastward drifting closed high pressure center. Cluster 5 is somewhat like cluster 1 of S7 with a trough being replaced by a high pressure center in the south, but in this case the high is already well established. Cluster 6 appears similar to Cluster 7 of S7, with a stagnant ridge and westerly flow in the north.

2) VARIATION OF TIME PERIOD

The clusters in S7 used distance vectors from 1979, 1981 and 1983 for ten sites. Clustering was also done using Ward's method on twenty sites (Alamo and Algoma omitted) for two 2.5 year halves, created by dividing 1979–1983 in half centered on 30 June 1981 (12-hour upwind distance vectors). These two halves can be compared to examine annual variability in type and frequency of clusters and can also be compared to the clusters obtained with ten sites and three years. This clustering yielded similar halves in terms of mean patterns and both halves were similar to the mean patterns seen in the S7 experiment. As shown in this and the previous section, the addition of more sites does not seem to have an appreciable effect on the cluster results.

3) VARIATION IN LOCATION OF INCLUDED SITES

Another sensitivity experiment involved retaining ten sites for clustering but changing their locations.

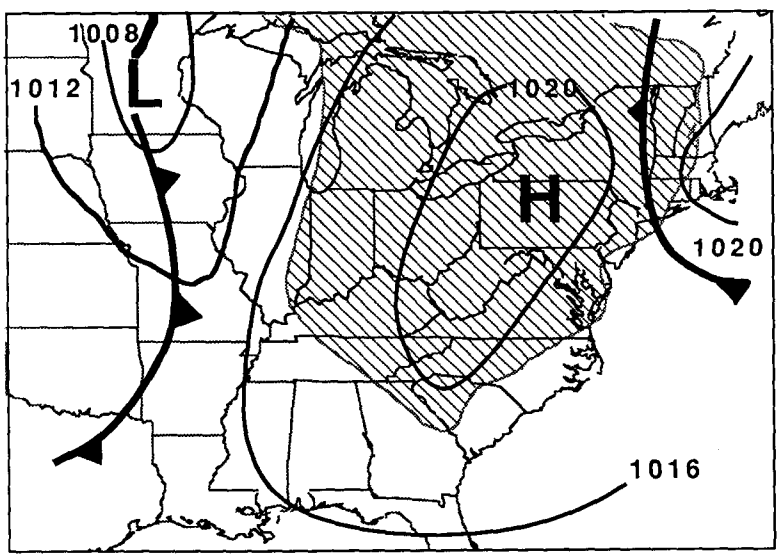

FIG. 12. Surface weather map for 0700 EST 14 June 1983. Area of haze is shaded.

The ten sites used in the S7 experiment were replaced by ten other sites while attempting to maintain the areal coverage and the distribution of the sites and keeping method, years and upwind distance the same. Hereinafter the first ten sites will be referred to as group A and the second ten as group B. Step 7 was examined in detail to provide comparison with group A. Based on maximum number of common events and comparison of the Euclidean distances among clusters for the mean transport vectors of group A versus those of group B, summed over site, the clusters from A could be matched one to one with the clusters from group A. Visual inspection of the mean patterns for the clusters confirmed the similarities. The best match occurred between the clusters representing the eastward moving anticyclone.

4) OTHER CLUSTERING METHODS

Using the 1979, 1981 and 1983 twelve-hour upwind distance vectors at the ten sites in Group A, clustering was done using some of the other methods available in addition to Ward's method. Used were centroid, median and average link, all with Euclidean distance, and average link with the correlation coefficient as the distance measure (Fox and Guire 1976; Anderberg 1973). Table 3 gives the cluster membership results at step 7 for the various methods. It is clear from the table that "chaining," as described by Anderberg (1973), is taking place in these methods, with one extremely large cluster and six smaller clusters representing outlier events being formed. This is in sharp contrast to Ward's method and its tendency to produce clusters of equal size.

The outlier clusters seem to mostly represent well-organized winter and spring storms with very distinct flow, based on a spot check of the Daily Weather Maps. If one's goal is to locate severe winter storms then these methods may be superior to Ward's method. However, the clusters appear to be too small to be of use in a more generalized classification or in examining precipitation chemistry. The results here differ from those of Kalkstein et al. (1987), who found that the average link method of joining clusters was better suited to their application than were the centroid or Ward's methods.

For the average link (AL) and average link with correlation coefficient as distance measure (ALCC) runs the disparity among cluster memberships was maintained at higher step sizes examined, such as steps 18 and 30. The mean transport of the large AL cluster resembles the S7 missing data pattern and is probably representative of average annual transport. Most of the other AL clusters are composed of closed lows and troughs with high wind speeds and have memberships too small to allow meaningful determination of their ability to distinguish pollutant classes. As noted before, average link appears best suited for classifying deep and organized storm systems. The ALCC clusters show little one-to-one mapping with the Ward's method results, as shown by the correspondence matrix between the Ward's step 18 group A clusters and the ALCC clusters as well as the median transport vector patterns. The merging into one ALCC cluster of Ward's clusters with similar patterns but differing wind speeds and the more uniform distribution of the ALCC clusters with time of year suggest that this may be due to the correlation coefficient being only sensitive to pattern while the Euclidean distance can separate similar patterns with different wind speeds. The median plots for the ALCC at step 18 generally resemble either mergers of several Ward's clusters, patterns similar to Ward's but displaced in location or time, or distinct but rare patterns not discriminated by the Ward's method. One cluster (10% of the data) resembles cluster 9A (step 18) with a "waffling high" moving north and south in the coastal states. Two clusters resemble clusters 4A and 5A at Ward's step 7.

5. Conclusions

Ward's method applied to variables representing the movement of air masses in eastern North America has produced a set of groups whose mean transport fields are significantly different from one another and correspond to actual weather patterns. Individual cases usually resemble the mean and demonstrate that the patterns are recurring. The clusters are relatively insensitive to changes in number of sites or years used and number of hours upwind used to define the distance vectors. A subsequent paper (Fernau and Samson 1990) will show that the clusters also have differing precipitation chemistry associated with them. The properties of different clustering methods lead to differences in the resulting clusters; Ward's method yields a relatively small number of clusters of similar size that

TABLE 3. Number of events in each cluster at step 7 for various clustering methods. 1979, 1981 and 1983 12-hour upwind data for group *A* sites. Avg. link, Corr: average link method with correlation coefficient used to measure distance.

	Cluster						
Method	1	2	3	4	5	6	7
Centroid	786	1	1	1	2	2	2
Median	765	1	2	24	1	1	1
Average link	723	1	14	35	1	19	2
Avg. link, Corr	512	162	43	72	3	2	1

depict the different types of flow patterns found over eastern North America while outliers in a dataset or extreme weather patterns are better revealed by average linkage or centroid methods. Currently being explored are ways to improve the clustering method through use of objectively analyzed data fields in place of trajectories, more years of data, and better methods for determining the optimum number of clusters.

Cluster analysis has been shown to be a useful tool in the computer-assisted classification of spatial patterns of weather variables and pollution data and should be considered for use along with more widely used synoptic climatological tools such as principal component analysis and correlation analysis.

Acknowledgments. This work was funded as part of the National Acid Precipitation Assessment Program by the United States Environmental Protection Agency. The results have not been subject to the agency's peer and policy review and therefore do not necessarily reflect the views of the agency, and no official endorsement should be inferred. Dr. Jennie L. Moody provided the initial idea for this work and invaluable advice along the way. Dr. John Merrill of Rhode Island got us back on the track of using transport to characterize similar meteorological regimes. Jeff Brook and Theresa Barnard of the University of Michigan helped with the analysis. The EPA project reviewers pointed out many useful revisions, corrections, and areas for further research.

REFERENCES

Anderberg, M. R., 1973: *Cluster Analysis for Applications.* Academic Press, 359 pp.

Anyadike, R. N. C., 1987: A multivariate classification and regionalization of West African climates. *J. Climatol.*, **7**, 157–164.

Barry, R. G., and A. H. Perry, 1973: *Synoptic Climatology—Methods and Applications.* Methuen & Co, 555 pp.

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 formulation. *J. Geophys. Res.*, **92**, 14 681–14 700.

Fernau, M. E., 1988: Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Ph.D. dissertation, University of Michigan 331 pp. [Available from University Microfilms International, Ann Arbor, MI.]

——, and P. J. Samson, 1990: 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.

Fox, D. J., and K. E. Guire, 1976: Documentation for MIDAS. Statistical Research Laboratory, The University of Michigan, 203 pp.

Gadgil, S., and N. V. Joshi, 1983: Climatic clusters of the Indian region. *J. Climatol.*, **3**, 47–63.

Galliani, G., and F. Filippini, 1985: Climatic clusters in a small area. *J. Climatol.*, **5**, 487–501.

Goossens, C., 1985: Principal component analysis of Mediterranean rainfall. *J. Climatol.*, **5**, 379–388.

Gordon, A. D., 1981: *Classification.* Chapman and Hall, 193 pp.

Heffter, J. L., 1980: Air Resources Laboratory atmospheric transport and dispersion model. NOAA Tech. Memo. ERL ARL-81, Air Resources Laboratory, NOAA, 24 pp.

Kahl, J. D., and P. J. Samson, 1986: Uncertainty in estimating boundary-layer transport during highly convective conditions. *J. Climate Appl. Meteor.*, **27**, 1024–1035.

Kalkstein, L. S., and P. Corrigan, 1986: A synoptic climatological approach for geographical analysis: Assessment of sulfur dioxide concentrations. *Ann. Assoc. Amer. Geogr.*, **76**, 381–395.

——, G. Tan and J. A. Skindlov, 1987: An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Climate Appl. Meteor.*, **26**, 717–730.

Lawson, M. P., R. C. Balling, Jr., A. J. Peters and D. C. Rundquist, 1981: Spatial analysis of secular temperature fluctuations. *J. Climatol.*, **1**, 325–332.

LeDrew, E. F., 1983: The dynamic climatology of the Beaufort to Laptev Sea sector of the Polar Basin for the summers of 1975 and 1976. *J. Climatol.*, **3**, 335–359.

——, 1985: The dynamic climatology of the Beaufort to Laptev Sea sector of the Polar Basin for the winters of 1975 and 1976. *J. Climatol.*, **5**, 253–272.

Maheras, P., 1984: Weather-type classification by factor analysis in the Thessaloniki area. *J. Climatol.*, **4**, 437–443.

Maryon, R. H., and A. M. Storey, 1985: A multivariate statistical model for forecasting anomalies of half-monthly mean surface pressure. *J. Climatol.*, **5**, 561–578.

Moody, J. L., and P. J. Samson, 1989: The influence of atmospheric transport on precipitation chemistry at two sites in the midwestern United States. *Atmos. Environ.*, **23**, 2117–2132.

Ronberg, B., and W.-C. Wang, 1987: Climate patterns derived from Chinese proxy precipitation records: An evaluation of the station networks and statistical techniques. *J. Climatol.*, **7**, 391–416.

Schulz, T. M., and P. J. Samson, 1988: Nonprecipitating low cloud frequencies for central North America: 1982. *J. Appl. Meteor.*, **27**, 427–440.

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

Wolter, K., 1987: The Southern Oscillation in surface circulation and climate over the tropical Atlantic, eastern Pacific, and Indian Oceans as captured by cluster analysis. *J. Climate Appl. Meteor.*, **26**, 540–558.